Uncovering rare genetic variants predisposing to coeliac disease

Vanisha Mistry

PhD Thesis
2013

Centre for Digestive Diseases
Blizard Institute
Barts and the London School of Medicine and Dentistry
Queen Mary University of London

Abstract

Coeliac disease is a common (1% prevalence) inflammatory disease of the small intestine, involving the role of tissue transglutaminase and HLA-DQ binding immuno-dominant wheat peptides. The disease is highly heritable, however, at most only 40% of this heritability is explained by HLA-DQ and risk variants from genome wide association and fine mapping studies. The hypothesis of the research in this thesis is that rare (minor allele frequency <0.5%) mutations of large effect size (odds ratios ~2 – 5) exist, especially in multiply affected pedigrees, which account for the missing heritability of disease.

NimbleGen exome capture and Illumina GAII*x* high throughput sequencing was performed in 75 coeliac disease individuals from 55 multiply affected families. Candidate genes were chosen from various analytical strategies: linkage, shared variants between multiple related subjects and gene burden tests for multiple potentially causal variants. Highly multiplexed amplicon sequencing, using Fluidigm technology, of all RefSeq exons from 24 candidate genes in 2,304 coeliac cases and 2,304 controls was performed to locate further rare variation. Gene burden tests on a highly stringent post quality control dataset identified no significant associations ($P<1x10^{-3}$) at the resequenced candidate genes.

The strategy of sequencing multiply affected families, and deep follow up of candidate genes, has not identified new disease risk mutations. Common variants (and other factors, e.g. environmental) may instead account for familial clustering in this common autoimmune disease.

Contents

Tables list

Chapter 5

Figures list

Chapter 1

Chapter 3

Chapter 4

Chapter 5

Chapter 7

Acknowledgments

The completion of the work in this PhD thesis would not have been possible without the following individuals whom I wish to extend thanks. Foremost I would like to express my sincere gratitude to my supervisor, Professor David van Heel, for continuous support of my PhD study, especially during my bioinformatic angst – your patience was highly appreciated. I also thank Dr Vincent Plagnol for his statistical input in my research. I would like to thank past and present members of the DvH group: to Dr Karen Hunt for your ongoing supportive words and research input, I am extremely lucky to have worked alongside you, to Nick Bockett for help with laboratory and sequencing techniques, and to Dr Graham Heap and Dr Patrick Dubois for setting great examples. I would also like to thank my fellow department and institute members, Professor David Kelsell, Dr Tania Marchbank, Nicola Kingston, Anna Alongi and Abdul Monaf, for always being there for guidance and advice when needed.

My family deserves my heartfelt thanks; I would not have made it here if not for your continuous backing, loving support and ambition for my scientific career, for this I am continually thankful and deeply appreciative. Finally, to my boyfriend Mel – your calmness has got me through a busy and stressful career step, I am deeply grateful and extremely lucky.

Statement of Originality

This statement outlines individuals contributing to the work carried out in this thesis project. Unless stated otherwise, all individuals have been acknowledged here and in publications.

DNA samples and preparation

DNA samples from multigenerational European and American families used for exome sequencing, genotyping and linkage analysis were provided by the following individuals:

- Paul Ciclitira at St Thomas' Hospital, London.
- Susan Neuhausen at Beckman Research Institute at the City of Hope, California.
- Åsa Naluai at Gothenburg University, Sweden.

David van Heel and I collected additional UK family samples via email/letter contact. Saliva DNA extractions and quantifications were performed by Dr Karen Hunt, Dr Patrick Dubois, Nick Bockett, Dr Graham Heap and myself. Dr Karen Hunt, Dr Patrick Dubois and Dr Graham Heap extracted DNA from blood samples (cases and controls) prior to my joining the group in 2008, which were also used for exome sequencing and genotyping. Exome control DNAs (222) from neurological disease samples were made available by Dr Vincent Plagnol at University College London Genetics Institute.

Exome capture, sequencing and data analysis

I performed all library preparations for NimbleGen exome array (60 individuals) and in-solution captures (75 individuals). Nick Bockett, Kerith Rae-Dias and I carried out Illumina GAIIx high throughput sequencing at Barts and the London Genome Centre. David van Heel generated all fastq files and Vincent Plagnol performed subsequent alignments and variant annotation for all exomes at University College London Genetics Institute. Vincent Plagnol, David van Heel

and I performed exome variant analysis. I performed all Sanger sequencing validation experiments.

Immunochip genotyping and linkage analysis

I ran all Immunochip genotyping chips for coeliax exome samples at Barts and the London Genome Centre and performed genotype data analysis. Vincent Plagnol performed linkage analysis at University College London Genetics Institute.

Candidate gene resequencing library preparation and data analysis

I performed all Fluidigm Access Array library preparations for 2,304 coeliac cases and 2,304 controls. I performed sequencing on the Illumina MiSeq at Barts and the London Genome Centre for all libraries. Muddassar Murza carried out high throughput sequencing for all libraries on the HiSeq 2000 at the NIHR GSTFT/KCL Biomedical Research Centre at Guy's Hospital. Michael Simpson generated fastq files for 4,608 samples. I performed sequence alignments, variant annotations and rare variant analysis on the entire dataset.

Abbreviations

λ_s	Sibling recurrence ratio
1000G	1000 Genomes Project
ATI	Amylase/trypsin inhibitors
CD	Coeliac disease
CDCV	Common disease common variant
CDRV	Common disease rare variant
CNV	Copy number variant
Encode	Encylopedia of DNA Elements
eQTL	Expression quantitative trait loci
GA	Genome Analyzer
GFD	Gluten free diet
GWAS	Genome wide association studies
HLA	Human leukocyte antigen
HWE	Hardy Weinberg equilibrium
IBD	Identical by descent
IBS	Identical by state
IEL	Intestinal intraepithelial lymphocytes
IFC	Integrated Fluidic Circuit
InBD	Inflammatory bowel disease
Indel	Insertion-deletion
LD	Linkage disequilibrium
LOD	Logarithm of odds score
LoF	Loss of function
MAF	Minor allele frequency
MHC	Major histocompatibility complex
NGS	Next generation sequencing
NHLBI	National Heart, Lung and Blood Institute
NPL	Non-parametric linkage
OR	Odds ratio

PCR	Polymerase chain reaction
qPCR	Quantitative PCR
RA	Rheumatoid arthritis
RFLP	Restriction fragment length polymorphism
SKAT	Sequence kernal association test
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
T1D	Type 1 diabetes
TDT	Transmission disequilibrium test
TG2	Tissue transglutaminase
WGA	Whole genome amplification
WGS	Whole genome sequencing
WTCCC	Wellcome Trust Case Control Consortium
UTR	Untranslated region

Bioinformatic resources and software

1000 genomes dataset, 2010 and 2012 data releases
www.1000genomes.org

Annovar annotation software
www.openbioinformatics.org/annovar

FastQC
www.bioinformatics.babraham.ac.uk/projects/fastqc

Genetic Power Calculator
http://pngu.mgh.harvard.edu/~purcell/gpc

Genome Analysis Toolkit (GATK) versions 2.3-9 and 2.4-7
www.broadinstitute.org/gatk

Merlin Linkage software
www.sph.umich.edu/csg/abecasis/merlin/tour/linkage

Novoalign mapping tool
www.novocraft.com

Phastcons
www.compgen.bscb.cornell.edu/phast

Picardtools
www.picard.sourceforge.net/index.shtml

PLINK
www.pngu.mgh.harvard.edu/~purcell/plink

PLINK/SEQ
www.atgu.mgh.harvard.edu/plinkseq

Polyphen and Polyphen-2
www.genetics.bwh.harvard.edu/pph2

R statistical package
www.r-project.org

Samtools versions 0.1.16 and 0.1.18
www.samtools.sourceforge.net/samtools

SeattleSeq
www.gvs.gs.washington.edu/SeattleSeqAnnotation

VCFtools
www.vcftools.sourceforge.net

Chapter 1

Introduction

The work in this chapter has been published in eLS (Wiley Online Library)

1.1 Coeliac disease: a complex autoimmune disorder

Gluten sensitive enteropathy, commonly known as coeliac disease (CD), is a complex autoimmune disorder of the small intestine that carries a strong genetic component and an equally strong environmental trigger of gluten. When a genetically susceptible individual ingests gluten, an immune response triggers gut inflammation resulting in acute morphological changes. The identification of the human leukocyte antigen (HLA) DQ gene variants and their role in CD has greatly contributed to our understanding of disease. A strong genetic component comes from an individual's HLA-DQ genotype for the risk allele coupled with the intake of gluten that is key in initiating an abnormal immune response in genetically at risk individuals (Abadie, Sollid et al. 2011). Symptoms are only reversed whilst maintaining a gluten free diet (GFD). Despite the causative trigger being non-self, CD is described as an autoimmune disease due to it possessing autoimmune components during disease manifestation i.e. the role of HLA-DQ2.5/8 in binding negatively charged gluten peptides eliciting a CD4+ T cell response (Djilali-Saiah, Schmitz et al. 1998), activation of intestinal intraepithelial lymphocytes (IELs) driving tissue damage (Chang, Mahadeva et al. 2005) and recent non-HLA associations that include an enrichment of genes predicted to control chemokine receptor activity, cytokine binding and production, and T and NK cell activation (Hunt, Zhernakova et al. 2008; Dubois, Trynka et al. 2010). Its genetic molecular basis can be described as a quantitative polygenic trait as the outcome phenotype is consequential of combinations of genes on multiple loci having an effect on each other.

1.2 Clinical manifestations and pathophysiology

Pathways for CD pathogenesis have been elucidated for many years; an atypical response to gluten produces an antigen-specific immunologic hypersensitivity in the mucosa of the small intestine leading to presentation of symptoms such as malabsorption, malnutrition, steatorrhea (diarrhoea caused by excess fat), weight loss, abdominal pain and anaemia. Samuel Gee was the first physician to

describe the disease in 1988, largely focusing on fat malabsorption. Consequentially, this led to the suggestion of removing fat from the diet to combat steatorrhea but this did not lead to any promising cure and other contributions in the field correctly implicated the effects of wheat flour (Dicke, HA et al. 1953; Losowsky 2008). Later, an increase in intestinal enzyme tissue transglutaminase (TG2) activity in coeliac patients was reported (Bruce at al. 1985) and its upregulation in the subepithelial lamina propria has been widely studied (Hansson, Ulfgren et al. 2002; Skovbjerg, Hansen et al. 2004). In 1997, it was discovered that TG2 is the antigen of the biomarker endomysial antibodies (EMA) (Dieterich, Ehnis et al. 1997). It is now understood that intestinal inflammation is triggered by an adaptive T-cell mediated immune response to gluten proteins found in dietary prolamins – plant storage proteins in grains such as wheat (gliadin), barley and rye. The TG2 enzyme catalyzes an aberrant deamidation (removal of the amide group) of specific glutamine residues from dietary wheat gliadins.

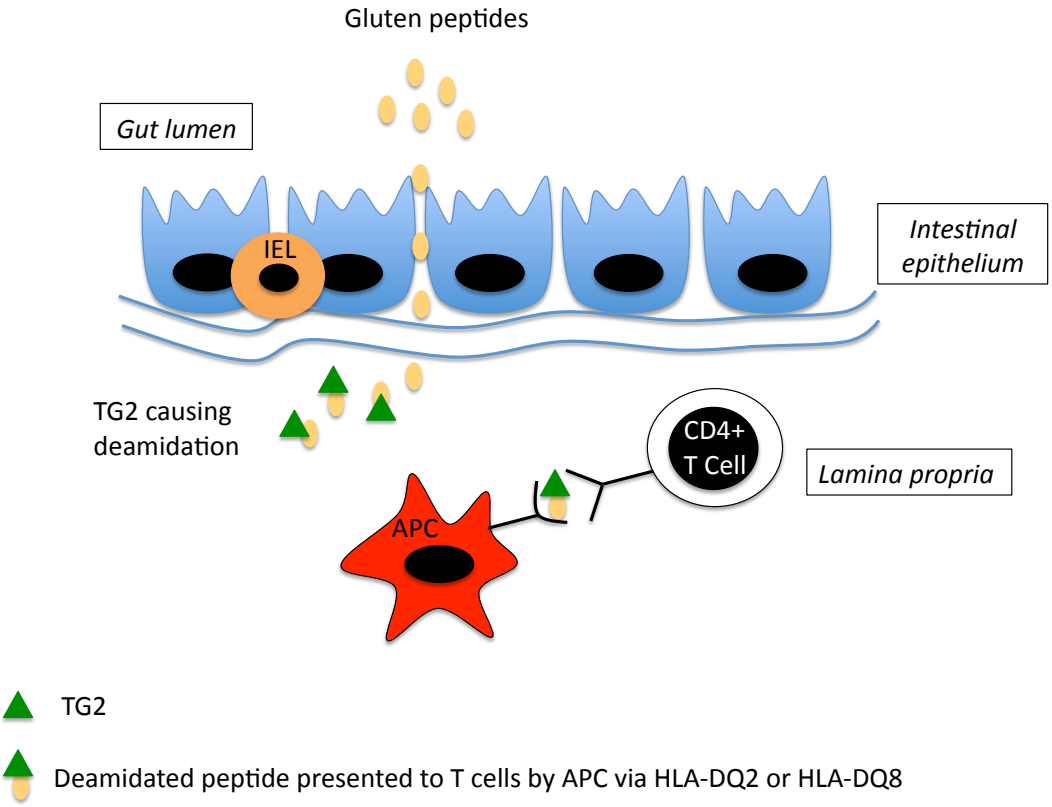
The process of the well-characterized adaptive response to gluten ingestion is illustrated in figure 1.1. The mucosa of the small intestine is covered by villi: finger-like projections that have a large surface area for absorption. Intestinal villous atrophy is when villi become truncated, inhibiting nutrient absorption (Figure 1.2). Other morphological changes characterizing CD enteropathy include hyperplastic crypts, epithelial cell damage and lymphoid infiltrates (Ferguson, Arranz et al. 1993). These changes occur as proline-rich areas of gliadin peptides are catalyzed to glutamic acids by TG2 (Sjostrom, Lundin et al. 1998); they pass through the epithelial barrier of the intestine into the lamina propria, instigating deamidation (Molberg, Mcadam et al. 1998). There is increased peptide affinity to HLA class II molecules (HLA-DQ2 or HLA-DQ8) generating CD4+ T-helper 1 cell (Th1) activation, including up-regulation of interferon (IFN)-γ production and T-bet levels in gut infiltrating cells (Holtmann and Neurath 2004).

Roles in the innate immune response have also come to light in recent years. The MHC class I receptor, *MICA*, is associated with the IEL (the majority of which are CD8+ T cells) response to gluten (Hue, Mention et al. 2004). IELs interact

with secretory enterocytes, which secrete *IL15* inducing stimulation and activation through *NKG2D* receptors producing a cytotoxic inflammatory response. Armed effector IELs are activated to lymphokine-activated killing cells under high doses of *IL2*, producing epithelial cell death in a T cell receptor – independent manner (Meresse, Chen et al. 2004). Additionally, a role for Th17 cells has also been suggested; these cells differ in their role in terms of cytokine production and a high expression of interferon regulatory factor-4 has been shown as a feature of gliadin-specific cells from CD patients (Castellanos-Rubio, Santin et al. 2009; Fernandez, Molina et al. 2011).

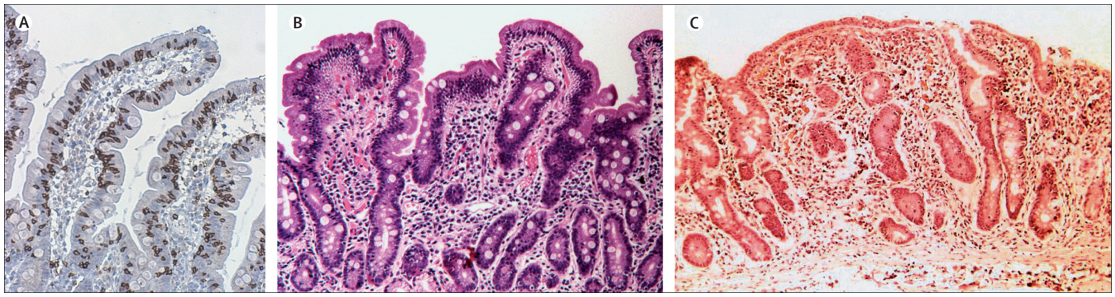
Other strong activators of the innate immune responses are monocytes, macrophages and dendritic cells. A recent study showed that gliadin fragments induce Th1 cytokine production by dendritic cells (Palova-Ielinkova et al. 2011) and amylase/trypsin inhibitors (ATI), which engage the toll-like receptor-MD2-CD14 complex, have been shown to release pro-inflammatory cytokines in cells. The activation of the *TLR4* signaling through ATI's may be a novel contributor to disease as mice deficient in *TLR4* are protected from intestinal immune responses upon ATI oral challenge (Junker, Zeissig et al. 2012).

Figure 1.1: Model of deamidated gluten peptide presentation by APC to T cells for subsequent loading onto HLA-DQ2 or HLA-DQ8 heterodimers



Adapted from Meresse, Ripoche et al. 2009 (Meresse, Ripoche et al. 2009). The removal of the amide functional group from an organic compound is described as **deamidation**. In CD, this chemical reaction degrades proteins as it damages the amide-containing side chains of glutamine.

Figure 1.2: Stages of villous atrophy in coeliac disease



Haematoxylin and eosin stain images of varying stages of instestinal villous atrophy: A - infiltrative non-atrophic lesions, B – atrophic lesions with shortened but detectable villi, C – complete villous atrophy. Taken from Di Sabatino and Corazza 2009 (Di Sabatino and Corazza 2009).

1.3 Disease epidemiology

While gluten is the only known environmental factor for CD occurrence, multiple inherited genetic factors affecting disease susceptibility have been postulated. Early evidence was highlighted by a dizygotic twin concordance rate of 20% when sharing two HLA haplotypes identical by descent, compared to a monozygotic pairwise concordance rate of 75%-80% (Greco, Romino et al. 2002; Di Sabatino and Corazza 2009). The sibling recurrence risk ratio (λ_s) is 10% for disease development (Greco, Romino et al. 2002; Polanco 2008). Approximately 1 in 100 individuals of white European descent (Lohi, Mustalahti et al. 2007) and 0.4 - 0.95% of individuals in the USA (Dube, Rostom et al. 2005) have CD, but it is less common in Asia and South America (Cummins and Roberts-Thomson 2009; Kotze 2009).

The dispersion of HLA-DQ variants in Europeans coincides with high disease prevalence in this population; HLA-DQ8 is most common in Northern Europe and Latin America, whereas HLA-DQ2 is most common in Western Europe as well as North and West Africa, the Middle East and Central Asia. High frequencies of HLA-DQ2 are also present in the Saharawi population of Algeria, where the prevalence of disease is 5.6% (Catassi, Ratsch et al. 1999), contrasting

to almost negligible prevalence in the Chinese-Japanese population (Cummins and Roberts-Thomson 2009). CD is notably more common in woman (between 2-3:1 men to woman ratio) however there are no gender differences at age of onset (Bai, Brar et al. 2005).

Disease diagnosis for CD has proven to be ambiguous, largely due to symptoms presenting similarly to other gut diseases, such as inflammatory bowel disease (InBD), and the distinction between gluten sensitivity and coeliac requires clarification prior to any disease diagnosis. At present, the ratio of diagnosed to undiagnosed is 1:7 (Heap and van Heel 2009), but testing for disease presence has improved with more sensitive and specific serological screenings. Prior to biopsy, testing for the presence of immunoglobulin A auto-antibodies to endomysium is a specific marker for the presence of CD (Ferguson, Arranz et al. 1993; Dieterich, Ehnis et al. 1997). If positive, confirmation is necessary by biopsy of the duodenal mucosa where the Marsh classification system is used for diagnosis according to small bowel pathology. If villous atrophy is not observed an IEL count is taken for early signs of disease manifestation (Ferguson, Arranz et al. 1993; Chang, Mahadeva et al. 2005). A high number if IEL's with normal villous morphology is the earliest pathological change following gluten challenge, however this may only be a sign of gluten sensitivity. Furthermore, HLA typing can be useful for exclusion in patients with equivocal histological finding but carries low specificity (Kaukinen, Partanen et al. 2002).

Currently, adhering to a GFD prevents disease relapse and symptoms often reduce within a few weeks. There are subsets of patients who do not respond to GFD (between 2-5%), mainly those diagnosed at the age of 50 or above (Tack, Verbeek et al. 2010). This is known as refractory CD. The leading cause of death for these patients is enteropathy-associated T cell lymphoma (Al-toma, Visser et al. 2007).

Although disease remission is maintained through a GFD in the majority of patients, this fails to completely reverse the histological changes in the intestinal mucosa (Lanzini, Lanzarotto et al. 2009). An alternative treatment in the form of a dietary supplement of peptidases (peptidases destroy proline and glutamine rich peptides) was taken to clinical trials in 2009 and provided

evidence of reduced immunological activity following a gluten meal, however no reduction of symptoms were noted (Tye-Din, Anderson et al. 2010). More recently, the use of human hookworm for the suppression of immune responses to gluten has been described in coeliac patients, showing a decrease of circulating T-reg cells, *IFN*-γ and *IL17A* post infection (McSorley, Gaze et al. 2011; Croese, Gaze et al. 2013).

1.4 The major histocompatibility complex and coeliac disease

The major histocompatibility complex (MHC) associated genes play a major role in the immune system. Its primary function is to bind pathogenic peptide fragments and display them on the cell surface for T-cell recognition. There is significant association with CD and a number of other autoimmune diseases, such as type 1 diabetes (T1D) (HLA-B and HLA-A class I genes) (Nejentsev, Howson et al. 2007), ulcerative colitis (HLA-DR2, HLA-DR9 and HLA-DRB1*0103 class II genes) (Stokkers et al. 1999) and ankylosing spondylitis, where polymorphisms in *ERAP1* affect risk of disease in HLA-B27 positive individuals (Evans et al. 2011). The MHC gene molecules are encoded on the short arm of chromosome 6p21. Coeliac associated MHC class II molecules are encoded at gene loci HLA-DQ, HLA-DP and HLA-DR (Sollid 2000).

1.4.1 The major histocompatibility complex

The MHC is one of the most hugely studied regions in the human genome due to its crucial role in immunity. The first MHC-encoded proteins were discovered on white blood cells and hence termed leukocyte antigens (commonly known as the HLA region). The MHC consists of 421 loci - 252 are expressed as genes, 30 are classified as transcripts and 139 are pseudogenes (Horton, Wilming et al. 2004). Early research in the MHC region confirmed presence of MHC-relevant genes extending beyond the defining boundaries of the region at that time, known as the classical MHC. The extended MHC was subsequently sequenced in

2003 and spans a 7.6Mb region on the short arm of chromosome 6 (Mungall, Palmer et al. 2003).

The MHC class I super cluster (a super cluster is defined as clusters with related genes outside the core cluster but within the extended MHC) are present on nearly all nucleated cells and include three classical class I genes (HLA-A, -B, -C). The genes encode a transmembrane heavy chain, which possess two polymorphic domains, α1 and α2, responsible for binding peptides. The bound peptides are presented to CD8+ cytotoxic T cells via the endogenous pathway. Apart from four non-classical genes and 12 pseudogenes, there are also class-I like genes such as the stress response genes *MICA* and *MICB*. The latter is located 47kb centromeric to HLA-B. Promoter polymorphisms in this gene highlighted an association with CD, with one out of the four promoter polymorphism haplotypes being overrepresented in CD patients (Rodriguez-Rodero, Rodrigo et al. 2006). Additionally, the expression profile of these class I-like MIC genes indicate a possible role in the mucosal immune system of the gut (Bahram 2000). The class II cluster includes HLA-DP, -DQ and –DR that encode α and β chains expressed as heterodimers on the cell surface and are responsible for antigen presentation to CD4+ T cells via the exogenous pathway (Watts 2004). Both class I and II cover a 3.6Mb region and in between both clusters is class III - this supercluster is the most gene dense of anywhere in the genome with 61 expressed genes (Xie et al. 2003).

Almost all autoimmune diseases show significant association with genes encoded in the MHC. For example, the tumor necrosis factor superfamily, a cluster of genes in the class III region, encodes proteins which play key roles in inflammation and immunity, and has been associated with susceptibility to autoimmune diseases such as T1D (Kumar, Goswami et al. 2007; Stayoussef, Benmansour et al. 2010) and neonatal lupus (Clancy, Marion et al. 2010). In spite of these associations, it's difficult to elucidate susceptibility genes in some diseases due to the complexity of linkage disequilibrium (LD) patterns and coinheritance between genetic polymorphisms in the region. The MHC sequencing consortium published an extensive map of LD patterns across the MHC (de Bakker, McVean et al. 2006) highlighting increased LD between

haplotype blocks resulting in ancestral haplotypes that were found to be common in Northern Europeans. LD was lower in Africans with shorter haplotypes. Also, high density gene clusters makes variant mapping complex - the two largest gene clusters in the extended MHC are of the histone and transfer RNA genes and it is thought these might be under selective pressure to cluster in order to maximize transcription levels in chromosomal regions such as the MHC, which is a transcriptional hotspot as well as a recombination hotspot (Mungall, Palmer et al. 2003).

1.4.2 HLA association with coeliac disease

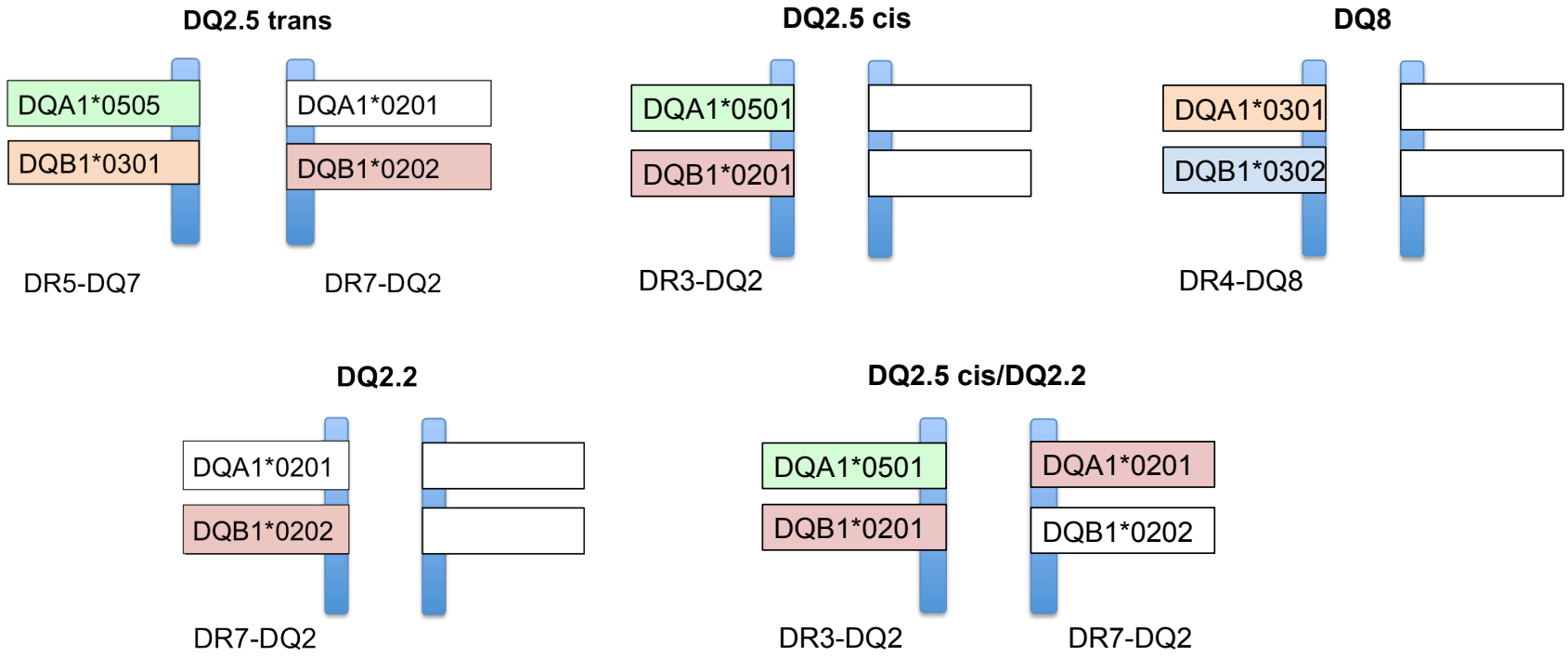
The most common genetic background coeliac individuals share is the presence of HLA class II genes HLA-DQ2 or HLA-DQ8 (Figure 1.3). HLA class II DQ2 molecules have a unique peptide-binding motif, which is necessary for CD, and encode for cell surface proteins on CD4+ lymphocytes that recognize gliadin peptides. These molecules are expressed on APCs (B cells, dendritic cells and macrophages). When the HLA-DQ2 serotype was identified as a true association it was found to be mediated through the DQ2.5 haplotype (Tosi, Vismara et al. 1983). The DQ2.2 haplotype does not appear to predispose to disease (Sollid, Markussen et al. 1989). HLA-DQ2 is encoded by HLA-DQA1*05 allele (alpha chain) and HLA-DQB1*02 allele (beta chain). The two alleles are present in *cis* conformation on the DR3 haplotype. 90% of European patients carry the HLA-DQ2 heterodimer and the remaining carries either one DQ2 allele or HLA-DQ8 (Karell, Louka et al. 2003). HLA-DQ8 is encoded by HLA-DQA1*03 (alpha chain) and HLA-DQB1*0302 alleles (beta chain).

It is possible to generate a combination of DQ2.2 and DQ2.5 haplotypes depending on parental genotypes, since each haplotype is only present on one chromosome. If in *cis* conformation, both alpha and beta chain are encoded on the same chromosome rather than each parent supplying one chain (*trans* conformation). One or two copies of HLA-DQ2 give an a priori intermediate or high risk for disease. Monsuur et al. (2008) used tagging single nucleotide polymorphisms (SNP) to predict HLA associated risk factors in CD. They found

that individuals with an increased disease risk were homozygous for the DQ2.5 haplotype or possessed a single copy of DQ2.5 and one copy of DQ2.2, DQ2.7 or DQ2.8 (Monsuur, de Bakker et al. 2008). This coincides with previous findings by van Belzen et al. who reported that being homozygote for DQ2.5 gives a 4-6 times increased risk of disease (van Belzen, Koeleman et al. 2004). Additional susceptibility coeliac alleles may be present in the MHC, however it is difficult to account for the high LD present in the region in a statistical test; any test outcome misses low frequency risk alleles due to the effect of common HLA-DQA1 and HLA-DQB1 high risk genotypes (Ahn, Ding et al. 2012).

The presence of HLA class II genes is not the sole genetic component for disease. HLA-DQ2 is expressed in 30% of the European population (Sollid, Markussen et al. 1989), with 2-5% of gene carriers developing disease. These early findings suggest other genetic factors contribute to the manifestation of CD.

Figure 1.3: HLA haplotype combinations in coeliac disease



White boxes denote 'other' haplotype. DQ2.5 *cis* is shown as a heterozygote; a DQ2.5 *cis* homozygote will carry same alleles on both chromosomes. Majority of CD patients express HLA-DQ2.5 encoded either in *cis* on the DR3-DQ2 haplotype, or in *trans* on the DR5-DQ7/DR7-DQ2 haplotype for heterozygous individuals. HLA-DQ2.2 confers low risk for CD if expressed solely. HLA-DQ8 is expressed in DQ2-negative patients (Abadie, Sollid et al. 2011). Adapted from Dubois and van Heel 2008 (Dubois and van Heel 2008).

1.5 Identifying susceptibility genes in complex disease

The past decade has been subject to a series of successful genetic studies all with the aim of finding risk variants susceptible to rare and common diseases. The identification of genetic factors for complex traits can be explained by efforts in Mendelian disease mapping. For these rare, monogenic traits, positional cloning was the traditional method to map a gene of interest using family based designs. Principal steps of this method are to identify and localize linkage to a small interval by performing successive rounds of mapping in families and then searching the area for mutations and relevance to disease. The idea is to find the gene and then assess its function in relevance to the phenotype and many Mendelian traits were mapped in this fashion. Difficulties in this approach in multifactorial traits (where phenotype is controlled by multiple genes) due to the weak relationship between genotype, at any given locus, and phenotype, led to the candidate gene approach. This strategy focused on identifying susceptibility variants through direct examination of biological candidates; for example, the *TNF* gene was shown to be important in the pathogenesis of a number of autoimmune and infectious diseases, such as malaria (Kwiatkowski, Hill et al. 1990) and an early candidate gene approach study found strong association with severe malaria of homozygosity for a *TNF* promoter SNP (Kwiatkowski, Hill et al. 1990). One gene belonging to the tumor necrosis factor family, *TNFAIP3I*, has been implicated in CD. This association was first found in 2010 by a more modern strategy, a genome wide association study (GWAS) (Dubois, Trynka et al. 2010). The application of GWAS was developed on the background of several years of linkage and candidate gene studies and largely commenced after extensive mapping of common and low frequency variation by the International HapMap Consortium. The following sections explain how SNPs have been mapped and exploited in GWA studies, the use of family-based association designs and linkage analysis to map chromosomal regions in related cases, and the strategy of GWA studies in complex disease using unrelated cases.

1.5.1 Single nucleotide polymorphisms and linkage disequilibrium

SNPs are the most common type of variation in the human genome and were the single most important finding from the Human Genome Project (HGP). Since then studies have assessed common and rare SNP frequencies in the human genome. A study in 2001 identified 3,899 SNPs in 82 unrelated individuals of varied ancestry, resulting in ~ 1 SNP every 185 bases present at a frequency of at least 5% in each population, and more polymorphic sites were observed in the 3' UTR regions (Stephens, Schneider et al. 2001). A recent study sequenced 202 genes in a larger sample set of 14,002 individuals and reported on average 1 SNP every 17 bases with a minor allele frequency (MAF) <0.5%, revealing an abundance of rare SNPs in coding regions (Nelson, Wegmann et al. 2012).

The differences in single nucleotides between two individuals contribute to phenotypic variability in human populations. For this reason they have been targeted for disease susceptibility in common traits. Almost all SNPs are biallelic and it's the differences in one of two alleles that are observed between individuals. Most SNPs are non-coding but two types of SNPs occur in coding regions of the genome: synonymous, which do not modify the amino acid sequence, and nonsynonymous, which change the amino acid sequence resulting in a deleterious effect on protein function. Nonsynonymous SNPs are further categorized into missense (substitution of an amino acid) and nonsense mutations (substitution to a premature stop codon in the mRNA transcribed sequence resulting in a truncated protein). Some missense mutations can also be deleterious and the most commonly known disease caused by such mutation is sickle cell anemia; an A→T mutation in the *HBB* gene causes an amino acid change in the β–globin protein. The mutation is pathogenic because a polar amino acid, glutamic acid, is replaced with a non-polar one, valine, causing adhesive interactions between heamoglobin molecules (Ingram 1957).

SNPs that are statistically associated are said to be on the same haplotype. A haplotype is a combination of DNA sequences at adjacent loci that are transmitted together. Nearly all variants result from a single historical mutation so that combinations of alleles at very close markers reflect ancestral

haplotypes. Therefore each new allele is initially associated with the other alleles that were present on the particular chromosomal background on which it arose, and this association is measured by the amount of LD i.e. alleles within a haplotype show LD and reside in recombination hotspots.

The concept of LD is centralized on the non-random association of alleles at different loci. Natural selection, or chance, caused the spread of common SNP mutations that arose thousands of generations ago. A second mutation occurring later but close to an earlier one results in both alleles being transmitted to the same offspring in subsequent generations. It is this model that is exploited in a GWAS (Xiong and Guo 1997). An increased risk of disease caused by one SNP denotes direct association between that SNP and disease in the population and indirect association between several nearby SNPs due to LD. Therefore it is possible to identify association in the chromosomal region without genotyping every SNP in a GWAS i.e. by using tagging SNPs. LD is prone to decay by recombination (since the probability of recombination increases with distance, the strength of LD between loci declines with distance) recurrence of the same mutation and gene conversion.

1.5.2 International HapMap Project, 1000 Genomes Project and Encode

The International HapMap Project commenced in 2002 with a focus to map all common genetic variation (greater than 5% MAF) across 11 populations (1,400 individuals), equating to 3.5 million SNPs. There have been 26 data releases so far capturing approximately 90% of genetic variation in the Caucasian population by using high throughput genotyping chips (Consortium 2003; Thorisson, Smith et al. 2005). This dataset was the first to describe the different types of variants, where they occur in our DNA and their distribution within and amongst populations. By comparing 1,400 individual DNA sequences, haplotypes could be deciphered by mapping chromosomal regions of shared genetic variants. This preceded the initiation and rise of many GWA studies as the HapMap provided a detailed measurement of genetic variation and LD patterns across major populations, as well as the identification of tag SNPs that

act as haplotype markers (Smith, Wang et al. 2006). Over the last decade the quantity of know variation has increased from 20% discovery by the HGP to 90% of mapped human variation with the help of HapMap and other similar projects. The 1000 Genomes Project (1000G) was set up in 2007 with a goal of identifying 95% of SNPs present at least 1% frequency in a range of populations (www.1000genomes.org). In the pilot phase, which commenced in 2008, three different strategies were used: high coverage sequencing of family trios to obtain true phasing of the variants detected, low coverage sequencing of many individuals (179) to allow broader detection of variants but requiring statistical phasing and sequencing of specific exon targets in a larger number of individuals (700) to allow detection of rare variants but would remain un-phased (Durbin, Abecasis et al. 2010). A main goal here was to reconstruct haplotypes using all variants typed from all datasets. The more recently published phase one dataset includes the genomes of 1,092 individuals from 14 populations (Abecasis, Auton et al. 2012). In this paper, functional variation was mapped by a combination of low coverage whole genome sequence data (2-6x read depth), targeted deep exome sequence data (50-100x), and dense SNP genotype data. The phase two dataset compiled in 2011 includes 1,715 individuals from 19 populations. The final phase three includes an additional 2,500 African and South Asian samples. This public reference catalogue of human genetic variation is already being used for imputation and will aid in identifying previously missed associations and provide a filter in Mendelian disease for exclusionary purposes.

Another project named Encylopedia of DNA Elements (Encode) published a myriad of papers in 2012 based on the identification of transcription regions, transcription factor association, chromatin structure and histone modifications in the human genome. This project differs completely from the genotype-based HapMap and 1000G projects and focuses on functional elements of gene products giving previously unknown insights into gene regulation and how statistical associations with disease correspond to these functional elements (Dunham, Kundaje et al. 2012).

1.5.3 Family based studies

Family based designs for the investigation of inherited disease have been used since Mendel's laws of inheritance dominated the fundamental concepts of genetics. Studies of extended pedigrees have several favourable features for novel gene discovery: causative gene pathways are more homogenous and there is a certain level of phenotypic control against genetic background and environmental exposures (Borecki and Province 2008). Gene mapping strategies utilize linkage and association studies, both of which use family data, but association studies can also be performed with unrelated individuals. A commonly used family based association test is the transmission disequilibrium test (TDT), first introduced in 1993 (Spielman, McGinnis et al. 1993). A TDT uses parents as controls for the cases, who are the affected offspring, so any confounding effects of population stratification are removed. The purpose of the test is to confer whether the disease allele is transmitted from parent to offspring more often in a disease population using genetic markers in nuclear families (trios) by mapping disequilibrium between the marker allele and disease locus. If the disease allele is transmitted to unrelated cases more often than expected by chance, this implicates a linked allele that is associated with the disease mutation. If the allele is only seen in related cases, then it becomes a test of linkage, not association. In essence, the TDT combines linkage and association approaches in cases where either performed separately have failed to provide a positive result. This test has been developed to include all family members and genotypic information (Abecasis, Cookson et al. 2000).

Where association analysis is powerful for the detection of common alleles that confer modest disease risk, linkage analysis is more powerful for identifying high-risk disease alleles. The independence of segregation, as inferred by Mendel's law of segregation, is not always true: there are group of traits which are linked and the genes controlling them tend to be inherited together by the offspring as a group, not independently. This is the underlying principle of a linkage study: if two individuals are phenotypically similar i.e. carry disease, then a genetic marker located near a disease susceptibility gene must also be

similar i.e. shared by both carriers. Linkage analysis searches for a high number of shared alleles than expected by chance amongst affected family members across regions of the genome. This indicates that there is a disease causing allele shared by affected individuals within the 'linked' region. Large regions of the genome are shared between closely related individuals inherited from the same common ancestor, so less than 500 polymorphic markers are usually sufficient to detect a linked region in an initial scan (Carlson, Eberle et al. 2004). The region of interest can only be narrowed down by further candidate-gene analysis because of the small number of recombination events within families. In the past, linkage maps were constructed using restriction fragment length polymorphisms (RFLP) as genetic markers and recombinant DNA technology (Botstein, White et al. 1980). On the back of that knowledge, the development of an RFLP linkage map was proposed to allow more powerful analytical strategies for the study of human diseases (Lander and Botstein 1986). Now, in silico approaches are used to map linkage in large multiplex families using several polymorphic markers for Mendelian and complex traits, such as autism spectrum disorders where great heterogeneity is observed (Szatmari, Paterson et al. 2007). The factors contributing to a linkage model, which will prove or disprove the null of no linkage, are the overall contribution of the trait loci and the genetic distance between the disease gene and marker being tested. Since this research project has a linkage component, this section is expanded on in the introduction of 'Chapter 4.2: Linkage Analysis with all Immunochip SNP variants'.

1.5.4 Heritability

Recent genetic studies have been attempting to close the gap on heritability estimates in common diseases and 'missing heritability' is a common phrase seen in most reviews surrounding this topic (Manolio, Collins et al. 2009; Eichler, Flint et al. 2010). Heritability is defined as the ratio of the genetic component to the total phenotypic variance. Main approaches to estimate heritability is based on the correlation of disease status in relative-pairs,

comparing disease concordances in twins, and using Bayesian methods to estimate genetic variance components in families of pedigrees (Tenesa and Haley 2013). Accurately testing heritability is important as knowledge of all genetic variants that account for disease can provide better insight into biological mechanisms. Estimating heritability has associated caveats, for example, assumptions of shared environment between twin pairs. However, this variation can be overlooked by ascertaining shared alleles identical by descent (IBD) by using genetic markers, making the estimate free from confounding non-genetic factors (Visscher, Medland et al. 2006).

GWAS attempt to explain heritability estimates by interpreting results in a population-specific manner, where a shared environment is assumed. In Manolio et al, evidence is stated that, for a GWAS, the premise is that a proportion of common diseases, most likely due to genetic variants, are heritable (Manolio, Collins et al. 2009). Visscher has published an array of papers describing software for the estimation of heritability from GWAS data; it estimates relatedness among samples, which is directly linked to the amount of phenotypic sharing. This multivariate approach was applied to seven diseases from the Wellcome Trust Case Control Consortium (WTCCC) study (Lee, Wray et al. 2011) and human height (Yang, Benyamin et al. 2010), where the estimate increased to 80%. However, this method used a set of all genotyped SNPs, including non-causal ones that mask correlation of causal SNPs, to estimate genetic correlation. Further optimization of this method based on a maximum likelihood estimation on causal SNPs increased accuracy (Golan and Rosset 2011).

GWA studies often report a large gap between population variance in disease and heritability estimates. Where a GWAS focuses on SNP frequencies, completely evaluating the overall genetic architecture of the genome for different traits can help in finding the missing genetic variance contributing to the total phenotypic variance in disease. Some examples of where missing heritability may lie are: epigenetic changes (Slatkin 2009), rare variants of relatively large effect, disease associated structural variants and copy number polymorphic duplications (Alkan, Kidd et al. 2009), epistasis (Zuk, Hechter et al.

2012), hundreds of more common variants and GWAS region tag variants underestimating the effect of correlated true causal variants. Furthermore, hidden gene-environment interactions and complex inheritance could also account for a substantial fraction of heritability.

1.5.5 Genome wide association studies

A GWAS using SNP markers is a more powerful approach for elucidating genetic determinants than family based linkage studies for a complex disease, due to its heterogenic nature and combined environmental effects. It was developed in tandem with the 'common disease common variant (CDCV)' hypothesis, recognizing that multiple genomic loci were likely to be involved in susceptibility to common multifactorial traits due to variants being present at relatively high frequency with an individually small magnitude of effect (Lander 1996; Risch and Merikangas 1996). With increasing sample size and wider coverage of the genome, more and more susceptibility loci for a wide range of complex diseases have been found.

The first GWA study was in 2005 and compared 96 subjects with age-related macular degeneration against 50 healthy controls (Haines, Hauser et al. 2005). Two years later, the WTCCC published the largest GWAS at the time, collecting subjects across seven common diseases, totaling 14,000 cases and 3,000 controls (Burton, Clayton et al. 2007). In 2009, over 500 GWAS studies in 300 diseases were published, of which more than 30 have been published in autoimmune disease (Baranzini 2009), and this number has risen sharply since then. Some studies have been follow-ups from linkage signals or candidate gene studies to narrow down association to a single haplotype, such as the region near the *CTLA4* gene in CD (Hunt, McGovern et al. 2005).

A GWAS uses SNP markers across the whole genome, which are tested for association with a disease in a large cohort of disease cases compared with a similar or higher number of controls. After performing correctional tests, common variants in correlation with disease are identified depending on the risk allele frequency, its association between marker genotyped and relative risk

conferred by genotype. For a successful GWAS large sample sizes, strict quality control, accurate genotyping to confer accurate phenotypes and careful adjustment for confounding factors are essential (de Bakker, Ferreira et al. 2008). Up to now associated variants have been found mostly in non-coding regions through GWAS, so it is accepted that common variant contribution to disease is more likely to be of regulatory function rather than protein coding.

Whilst GWAS has increased the number of identifiable disease-associated loci, the odds ratio (OR; disease risk if in possession of disease associated SNPs) for these risk variants have been fairly modest (commonly between 1.05-1.5), so the proportion of the risk attributable to genetic variants remains small. That said, there are other aims in a GWAS, which is to use the information in order to understand downstream processes leading to the observed disease phenotype. Linking the measured genotype to phenotype is often a painstaking task as in some cases the outcome genotypic association may have no reflection on the hypothesized disease pathway. Combining transcriptional network analysis and gene expression is required to discover regulatory gene networks (Keller, Martini et al. 2012).

When a GWAS has failed to find a significant finding, or the finding has not reached a significant statistical association, a meta-analysis is an excellent tool in combining GWAS results for the same phenotype in order to gain a more meaningful association. A meta-analysis automatically increases the statistical power due to a larger sample size, and may provide further support for known risk signals, as in T1D for the *IL2-IL21* region (Cooper, Smyth et al. 2008) or highlight previously unknown associations, as in rheumatoid arthritis (RA) where seven new risk alleles were identified in a sample size of ~42, 000 (Plenge, Stahl et al. 2010). Additional power can be achieved by combining phenotypes where there is a clear shared genetic foundation to identify shared risk alleles. A meta-analysis combining CD and RA published GWAS results identified four additional gene loci not previously confirmed in either disease, and also implicated four gene loci previously established in CD and RA to be significant in other autoimmune diseases (*SH2B3*, 8q24, *STAT4*, and *TRAF1-C5*) (Zhernakova, Stahl et al. 2011). Other successful studies have demonstrated the

power of meta-analyses and the advantages of collaborating with other GWA study cohorts in order to gain valuable insights into disease pathways (Barrett, Hansoul et al. 2008; Zeggini, Scott et al. 2008). Furthermore, genome wide imputation from a reference panel, such as HapMap or 1000G, can locate stronger associations. This has been highlighted for the *TAGAP* risk locus in RA (Plenge, Stahl et al. 2010; Chen, Stahl et al. 2011). Imputation was designed to allow testing of un-typed variants by combining SNP correlation patterns from a reference panel with genotype data on tagged SNPs (Servin and Stephens 2007). The genotype estimation is then tested for phenotype association and further assessed in a replication cohort. This method increases the statistical power to detect novel associations as well as identifying missed associated variants, for example, imputation from the 1000 genomes panel observed two variants (*IL2RA* associated with T1D and *CDKN2B* associated with type 2 diabetes) that were previously undetected in the WTCCC case control study (Burton, Clayton et al. 2007; Huang, Ellinghaus et al. 2012). 1000G-based imputation has proven to be an excellent tool in increasing genome-wide coverage in different populations because of a higher number of common and low frequency SNPs resulting in more accurate genotyping (Gao, Haritunians et al. 2012; Sung, Gu et al. 2012).

1.6 Known genetic structure of coeliac disease

As noted, the most significant association to CD so far is with HLA-DQ2. Possession of HLA-DQ2 serotypes is necessary for affinity to deamidated gliadin, yet 30% of the Caucasian population also carry HLA-DQ2 without developing disease (Heap and van Heel 2009). The following sections explain methods used to find other CD associations and their contribution to disease risk.

1.6.1 Regions identified through linkage analysis

To determine absolute risk of disease, non-HLA risk alleles in CD must be taken into account. Unlike Mendelian disease, complex disease has had less success in finding causal variants through linkage. In CD, linkage was found to various regions in early studies including 5q (Greco, Corazza et al. 1998; Greco, Babron et al. 2001; Percopo, Babron et al. 2003), which was replicated in a meta analysis of multiple populations (Babron, Nilsson et al. 2003), and 19p (Van Belzen, Meijer et al. 2003). Linkage to 2q33, containing *CTLA4*, *ICOS* and *CD28* which is involved in immune suppression, suggested several independent loci contributing to disease (Amundsen, Naluai et al. 2004). Haplotype analysis in this region showed strong association in the Irish population (Brophy, Ryan et al. 2006) as well as linkage in several other populations (Djilali-Saiah, Schmitz et al. 1998; Naluai, Nilsson et al. 2000; King, Moodie et al. 2002) and variants in the 3' region of *CTLA4* were thought to influence responses in T1D (King, Yiannakou et al. 2000). In spite of promising initial analysis, replication in genome wide scans was inconsistent for this region (King, Moodie et al. 2003).

A whole genome linkage approach followed by Immunochip genotyping in two multiply affected Finnish and Hungarian coeliac families found linkage (logarithm of odds score (LOD) >1.3) at 4q, 6p, 6q, 7p, 17p, 17q and 22p, but only variants at 4q, harbouring the *IL2-IL21-TENR* locus, segregated with disease in both families. This risk haplotype was estimated to have a 2% frequency estimation in the CEU population indicating that the observed linkage was due to a rare risk haplotype in this region not tagged by previously described common variants (Einarsdottir et al. 2012). This illustrates the need for genotyping chips with less than 5% frequency SNPs to search for those rare risk alleles with large effect size, which is discussed in more detail in section 1.7.

It is known that the power of family studies is decreased due to small effect sizes attributable to genetic variants present at high frequency (Kruglyak 2008). An exception is *NOD2* in Crohn's disease (Hugot, Chamaillard et al. 2001; Ogura, Bonen et al. 2001), and HLA replication in CD, due to common encoding variants being of large effect size hence having sufficient statistical power. This explains

that predisposition to complex disease is not caused by just a handful of highly penetrant mutations but a mixture of multiple risk variants with varying effect size.

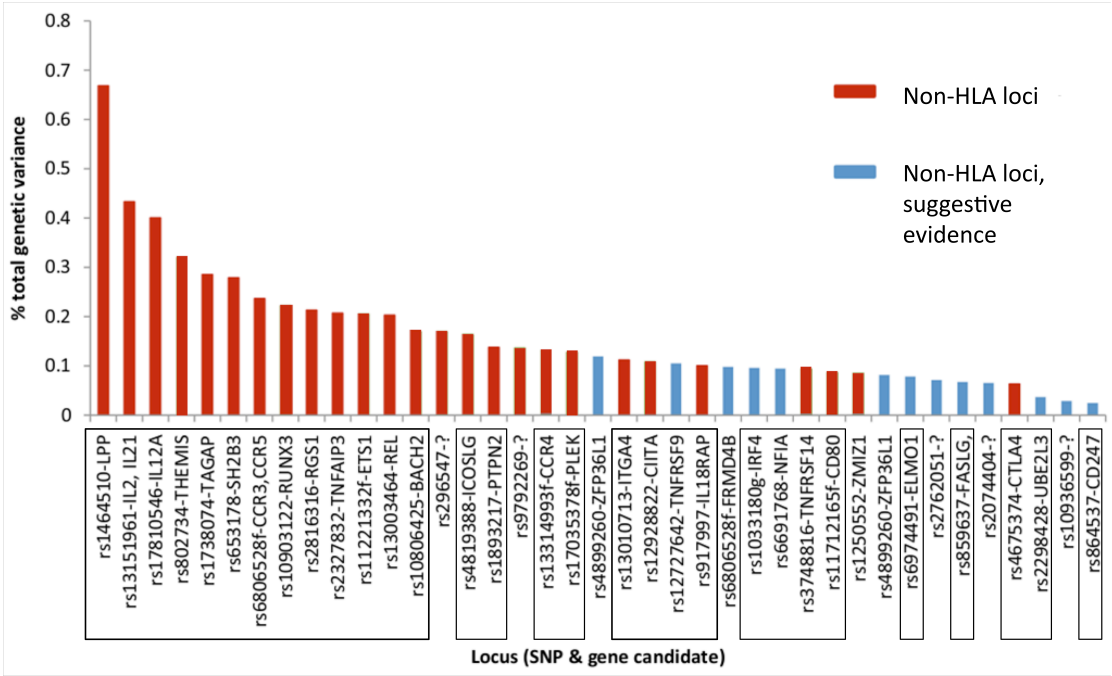
1.6.2 Susceptibility gene loci identified by GWAS and further dense genotyping with Immunochip

To date two GWA studies using samples of European ancestry have been carried out in CD identifying non-HLA variants. In the first GWAS, 778 coeliac cases and 1,422 matched population controls using 310,605 tag SNPs showed the highest association on chromosome 4q27 harbouring the *KIAA1109-TENR-IL2-IL21* LD block (van Heel, Franke et al. 2007). Follow up studies found associations in *REL*, *TNFAIP3* and a region encompassing *CTLA4*, *ICOS* and *CD28* (Smyth, Plagnol et al. 2008; Trynka, Zhernakova et al. 2009). The UK follow up replication study by Hunt et al. identified a further seven regions reaching genome-wide combined significance ($P \leq 5 \times 10^{-7}$) (Hunt, Zhernakova et al. 2008). An additional association was found in *ITGA4* in a US case control collection (Garner, Murray et al. 2009). Six out of the eight coeliac regions found in van Heel et al. (2007) and Hunt et al. (2008) were replicated in an Italian cohort (Romanos, Barisani et al. 2009).

The second generation GWAS by Dubois et al. used a larger sample size (discovery dataset included 4,533 European cases and 10,750 matched controls) and identified a further 13 genome wide significant regions, with evidence for an additional 13 suggestive loci upon replication, 28 of which contained genes controlling immune responses (Figure 1.4) (Dubois, Trynka et al. 2010). A study by Amundsen et al. used family TDT (to remove any sources of false positives) and found significant associations in four out of the nine CD regions tested in a Swedish-Norwegian family cohort (Amundsen, Rundberg et al. 2010). A meta-analysis combining both CD published GWAS datasets (Zhernakova, Stahl et al. 2011) with an RA sample cohort including 5,539 cases identified four novel SNPs for both diseases in *DDX6*, *CD247*, *UBE2L3* and *UBASH3A*. The SNP in *CD247* (rs864537) was previously identified in CD (Dubois, Trynka et al. 2010), but

reached a higher statistical significance in this larger meta-analysis dataset (P=2 x 10^{-11}). The other three new overlapping associations for CD and RA were later statistically significant in a CD fine mapping study carried out in 2011 (discussed below).

Figure 1.4: Total genetic variance contributed to CD by significant and suggestive 39 2010 non-HLA loci



HLA-DQ2/DQ8 is present in 45.3% of the population and 97.4% of coeliac individuals. Boxes surround immune genes. Adapted from Dubois et al. 2010 (Dubois, Trynka et al. 2010).

Fine mapping is a necessary step after genotyping in an attempt to refine associated region(s) to a causal variant(s) by analyzing a high density of genetic markers across the associated LD region. Using T1D as an example, extensive mapping across the MHC region for discovery of HLA-linked loci to T1D established HLA-B and HLA-A to be associated independently of HLA class II genes (Brown, Pierce et al. 2009; He, Hamon et al. 2009; Howson, Walker et al. 2009). Results from this study illustrated the multilocus effects due to classical

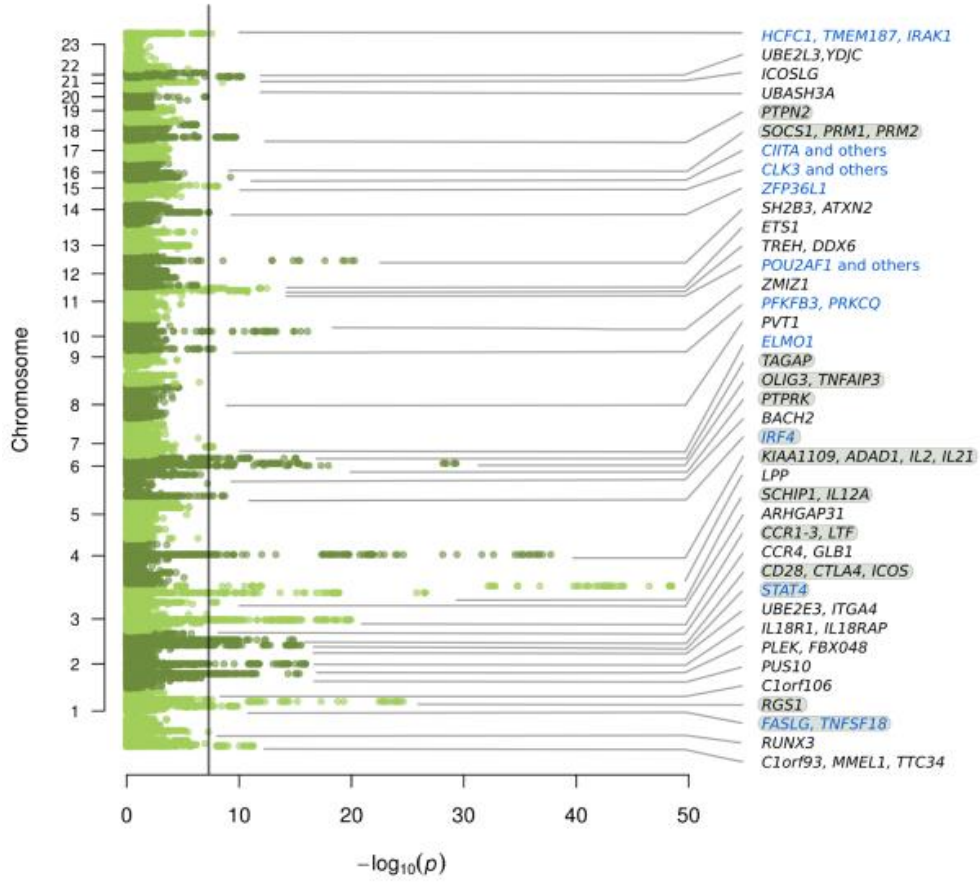
HLA genes and extensive LD spanning the entire region. Additionally, the major susceptibility gene to T1D was refined to two independent groups of SNPs encompassing *IL2RA* intron 1 and 5' regions of *IL2RA* and *RBM17* after large scale fine mapping (Lowe, Cooper et al. 2007). *IL2RA* is involved in the regulation of T cells correlating with functional work in a non-diabetic obese mouse model showing reduced expression of *IL2* relating to increased disease risk and reduced T-reg cell activity (Yamanouchi, Rainbow et al. 2007). A coeliac-associated locus, *IL2/IL21,* is also implicated in SLE and present research has localized the effect of this locus to two SNPs in high LD through fine mapping of 45 tag SNPs in the region (Hughes, Kim-Howard et al. 2011). Again, it is easy to identify multiple associated loci across autoimmune diseases by looking deeper into GWAS associated regions. Recently, this has been greatly expanded by the development of the Immunochip genotyping array.

In 2010 a deep replication effort to fine map autoimmune diseases was created by the Immunochip Consortium, providing an opportunity to refine GWAS signals and to identify new disease associations with loci implicated in other autoimmune diseases. In collaboration with Illumina, this Infinium chip contained 186 non-HLA risk loci including many more markers in both regulatory and exonic gene regions and all 1000G pilot CEU low frequency variants for dense SNP mapping analysis. The first paper to be published was in CD: a scan was performed in 12,041 coeliac cases and 12,228 controls of European and Indian origin. This study replicated previously described associations (Hunt, Zhernakova et al. 2008; Dubois, Trynka et al. 2010) and identified 13 new disease risk loci reaching genome wide significance (Figure 1.5) (Trynka, Hunt et al. 2011). In addition, over a third of loci containing multiple independent associations were a combination of common, low frequency and rare variants.

This study illustrates the effectiveness of the Immunochip array in refining risk associations, as 53% of signals were localized to a single gene. Despite this major advance, only 13.7% of genetic variance was found to be explained by 39 non-HLA associated risk loci in individuals of European ancestry (this increases to 40% with the HLA). Similar statistics are observed for other autoimmune

diseases: Eyre et al. identified 14 new RA loci at $P < 5x10^{-8}$ with Immunochip genotypes and seven additional loci when combined with GWAS meta-analysis data, totaling 51% of the heritability estimate, of which 36% is explained by the HLA (Eyre et al. 2012); 39 independent signals in psoriasis at $P < 5x10^{-8}$ account for 14.3% total variance or ~22% heritability (Tsoi et al. 2012); 26 independent SNPs and four HLA types account for 4.9% and 1.4% variance in liability, respectively, with the total heritability estimate at 16.2% in primary biliary cirrhosis (Liu, Almarri et al. 2012). Although heritability estimates have not entirely been explained, the Immunochip consortium has allowed progression in many immune-mediated disease types using different analytical strategies. For example, a massive InBD meta-analysis combining imputation-based association analysis using 15 GWAS datasets from Crohn's disease and ulcerative colitis and validation using Immunochip genotypes identified 193 independent association signals in 163 regions, of which 71 were new. Out of 163 loci, 110 were associated with both Crohn's and ulcerative colitis phenotypes, and 113 of these InBD loci were shared with other complex diseases (Jostins, Ripke et al. 2012). This study also highlighted shared responses in the host microbe through gene expression analysis.

Figure 1.5: Manhattan plot showing previously associated and new CD risk loci with significant threshold set at $P \leq 5x10^{-8}$



Novel loci are in blue and multiple signal loci are highlighted in grey. Taken from Trynka et al. 2011.

1.6.3 Overlap with other autoimmune diseases

The numbers of autoimmune disease loci that overlap with CD highlight shared immunological pathways e.g. *CTLA4* (T cell co-stimulation), *IL2-IL21* (B cell and T CD8 differentation), *ICOSLG* (B cell help*)*, *PTPN2* (IFNy and *IL76* signaling), and *IL18RAP* (signal transduction) (Meresse, Malamut et al. 2012). Approximately 64% of 39 known CD loci are shared with at least one other autoimmune disease (Gutierrez-Achury, de Almeida et al. 2011). Table 1.1 outlines 39 non-HLA coeliac risk loci and the GWAS associated overlapping autoimmune disease at $P < 5 \times 10^{-5}$. The reported gene function is also noted in this table,

highlighting processes involving T-lymphocyte proliferation, cytokine processes and signaling. The SNP with the largest significant association in CD is rs2030519 in *LPP* ($p=3.0 \times 10^{-49}$); this association is not shared in any of the obvious gut related immune-mediated diseases but is reported in vitiligo (Jin, Birlea et al. 2010), an autoimmune disease characterized by loss of pigmentation in skin and hair. This is unsurprising since the ubiquitous nature of the immune system can lead to alterations manifesting in other etiologically related diseases. Crohn's disease shares most loci with CD, but out of all of them *TAGAP* and *PUS10* have been firmly identified as shared risk loci after replication with combined *p* values of 1.55×10^{-10} and 1.38×10^{-11} respectively (Festen, Goyette et al. 2011). From previous meta-analysis and Immunochip studies in autoimmune diseases it is clear that one dataset combining different diseases yields more power to finding shared loci, however this can be further improved by proper phenotypic characterization as CD and other autoimmune diseases can manifest into a broad range of overlapping symptoms. Apart from clinical perspectives, the advantages of being able to deduce a shared immunological profile from information on loci effects in different diseases can allow investigation into a shared host microbiome, as observed between InBD loci and responses to mycobacteria (Jostins, Ripke et al. 2012) and interpretation of any protective associations. A study has shown that three out of the five risk markers in the *IL2/IL21* region showing strong association with CD have a protective effect in ulcerative colitis (Glas, Stallhofer et al. 2009) supporting the notion that alleles are in balancing selection if they increase risk of one disease but decrease the risk of another (Blekhman, Man et al. 2008).

Table 1.1: 39 non-HLA coeliac loci from Immunochip study (2011) showing association with other autoimmune diseases

Associated coeliac loci	Reported genes	Overlapping autoimmune diseases
1q24.3	*FASLG, TNFSF18*	Crohn's disease
1q32.1	C1orf106	Ulcerative colitis, Type 1 diabetes, Crohn's disease, Multiple sclerosis, Ankylosing spondylitis
1q31.2	*RGS1*	Multiple sclerosis, Type 1 diabetes
1p36.11	*RUNX3*	Psoriasis
1p36.32	*TNFRSF14, MMEL1*	Rheumatoid arthritis, Ulcerative colitis, Multiple sclerosis, Graves' disease
2q12.1	*IL18RAP, IL18R1*	Crohn's disease, Type 1 diabetes
2p14	*PLEK, FBX048*	Rheumatoid arthritis, Crohn's disease,
2p16.1	*PUS10*	Ulcerative colitis, Crohn's disease, Psoriasis, Rheumatoid arthritis
2q31.3	*ITGA4, UBE2E3*	Ankylosing spondylitis
2q32.3	*STAT4*	Crohn's disease
2q33.2	*CTLA4, ICOS, CD28*	Type 1 diabetes, Rheumatoid arthritis, Graves' disease
3q13.33	*ARHGAP31*	Vitiligo, Multiple Sclerosis
3p21.31	*CCR1-3, LTF*	Ulcerative colitis, Crohn's disease
3p22.3	*CCR4, GBL1*	None

3q25.33	*IL12A, SCHIP1*	Multiple Sclerosis
3q28	*LPP*	Vitiligo
4q27	*IL2, IL21, KIAA1109, TENR, ADAD1*	Ulcerative colitis, Type 1 diabetes, Rheumatoid arthritis, Crohn's disease, Psoriasis
6q15	*BACH2*	Type 1 diabetes, Crohn's disease, Multiple Sclerosis, Graves' disease
6q22.33	*PTPRK*	Crohn's disease, Multiple Sclerosis
6q23.3	*TNFAIP3, OLIG3*	Psoriasis, Rheumatoid arthritis, Systemic lupus erythematosus, Ulcerative colitis, Multiple Sclerosis, Type 1 diabetes
6q25.3	*TAGAP*	Crohn's disease, Multiple Sclerosis, Type 1 diabetes, Psoriasis
6p25.3	*IRF4*	None
7p14.1	*ELMO1*	Rheumatoid arthritis
8q24.21	PVT1	Multiple Sclerosis, Crohn's disease, Rheumatoid arthritis
11q23.1	*POU2AF1*	None
11q23.3	*TREH, DDX6*	Crohn's disease, Vitiligo, Multiple Sclerosis, Systemic lupus erythematosus, Rheumatoid arthritis
11q24.3	*ETS1*	Crohn's disease, Psoriasis
10p15.1	*PFKFB3, PRKCQ*	Multiple Sclerosis, Crohn's disease, Type 1 diabetes, Rheumatoid arthritis, Vitiligo

10q22.3	*ZMIZ1*	Multiple Sclerosis, Crohn's disease
12q24.12	*SH2B3, ATXN2*	Rheumatoid arthritis, Type 1 diabetes, Vitiligo
14q24.1	*ZFP36L1*	Type 1 diabetes, Crohn's disease, Multiple Sclerosis
15q24.1	*CLK3*	None
16p13.13	*CIITA, SOCS1, PRM1, PRM2*	Multiple Sclerosis, Type 1 diabetes, Ulcerative colitis, Crohn's disease, Psoriasis
18p11.21	*PTPN2*	Type 1 diabetes, Crohn's disease
21q22.3	*UBASH3A* *ICOSLG*	Type 1 diabetes, Crohn's disease, Rheumatoid arthritis, Ulcerative colitis
22q11.21	*UBE2L3, YDJC*	Crohn's disease, Psoriasis
Xq28	*HCFC1, TMEM187, IRAK1*	Type 1 diabetes

Coeliac loci with genome wide significance at $P < 5 \times 10^{-8}$; autoimmune disease overlapping loci with genome wide significance $P < 5 \times 10^{-5}$
Autoimmune diseases with overlapping loci according to the Catalogue of Published Genome-Wide Association Studies (http://www.genome.gov/26525384, refer to website for references. - Accessed on 17th February 2013) and published papers.

1.7 Finding further genetic causal variants in complex disease

GWA studies in CD and other complex diseases have been useful tools for expanding the genetic understanding of disease by identifying new loci and replicating previously associated loci through further fine mapping (Trynka, Hunt et al. 2011; Eyre, Bowes et al. 2012; Tsoi, Spain et al. 2012). Overall, GWAS findings have implicated numerous associated variants that are mostly common, have modest to weak effect sizes, and are credible disease markers, however only explain a small fraction of heritability. By design, GWAS focus on common variants (MAF >5%) to tag large haplotype blocks and overlook potentially functionally detrimental variation contributed by low frequency and rare/novel mutation. In order to expand the genetic architecture of disease and close the heritability gap, searching away from common ancestral haplotypes and looking at variation arising from recent, more rare genomic events is necessary.

Next generation sequencing (NGS) allows deeper visualization of types of variants not typically seen in genome wide scans of non-coding regions of the genome. In the past, sequencing studies have commonly been dismissed due to the associated high costs, but vast improvements since 2008 (the year sequencing transitioned from Sanger and capillary based to next, or second, generation sequencing) has reduced the cost from $10m per genome to $7,500 per genome in 2012 (Wetterstrand). Sequencing studies are now commonplace in genetic research (Metzker 2010) and are being used as a diagnostic tool for disorders observed in clinic, such as rare pediatric disorders (Kingsmore, Dinwiddie et al. 2011). Diagnosing rare conditions at birth often involves lengthy and costly investigations to establish disease etiology, so NGS combined with the advanced bioinformatic expertise has offered a breakthrough in prenatal diagnosis, allowing more targeted and systematic individual therapy from birth (Talkowski, Ordulu et al. 2012).

All commercially available NGS protocols have shared attributes: fragmentation of genomic DNA, ligation with custom adapters to create a DNA library and then library amplification on a solid surface (either a bead or glass). There is direct

step-by-step detection of each nucleotide base incorporated during the sequencing reaction and thousands of reactions are imaged per instrument per run, giving it the term 'massively parallel sequencing'. The read lengths are short compared to capillary sequencers and reads can be run single or paired end, meaning the linear DNA fragment is sequenced at both ends in two separate reactions i.e. two sequences from a single DNA fragment. Once DNA has been sequenced, it is aligned to a reference sequence panel. This is where paired end sequencing provides more accuracy when mapping reads in large and complex genomes but it has a higher cost and can be time consuming. The main companies providing NGS worldwide are Illumina, Life Technologies and Roche and all have their own associated benefits and drawbacks, the main ones being errors in calling the alternate allele, errors in variant calling and coverage differences due to G-C content (Shendure and Ji 2008; Ratan, Miller et al. 2013).

The question now is how to apply NGS in a research study to find novel genes in complex traits in which genetic heritability still remains largely unknown? The large-scale GWAS design has established many common associated disease risk variants with consistently low to modest effect sizes and an overall modest combined genetic variance, and further genotyping in larger sample sizes (>10,000) may achieve the power needed to increase the genetic variance attributed to common regulatory variants. Conversely, a spectrum of low frequency rare variants with functional effects on the transcribed protein and an intermediate to large effect size (OR of 2.5 and above) may possibly be the key to establishing more disease causing loci (Manolio, Collins et al. 2009). Low frequency (0.5% to 5% MAF) and rare variant (below 0.5%) analysis has been shown to contribute significantly to genetic architecture of disease (Coventry, Bull-Otterson et al. 2010; Durbin, Abecasis et al. 2010). Evidence suggests that an excess of rare variants are to be found in disease-associated genes and a recent study analyzing 6,515 exomes (protein-coding regions) predicted that most deleterious single nucleotide variants (SNV) only arose in the last 5,000 - 10,000 years and that disease genes contained more of these variants than other genes, highlighting implications for the prioritization of these genes in complex disease research studies (Fu, O'Connor et al. 2013). It has been

frequently proposed that rare mutations of large effect size account for a substantial proportion of the missing heritability in disease (Pritchard 2001; Eichler, Flint et al. 2010), coinciding with the 'common disease rare variant (CDRV)' hypothesis. This hypothesis is founded upon the fact that genes with loss of function (LoF) variants are sparsely observed in healthy individuals – those that are found are common at minor allele frequencies of >5% and in a very small number of nonessential genes, such as olfactory receptors, and do not result in any severe phenotype (Carlson, Eberle et al. 2004). Studies suggest that while genes with common LoF variants are likely to be benign, genes with low frequency LoF variants are more likely to be deleterious, simply because selection prevents these deleterious variants from reaching moderate to high allele frequencies (MacArthur, Balasubramanian et al. 2012). The following sections explain how targeted gene resequencing, exome and whole genome sequencing can be used in the search for these types of disease susceptibility variants in complex disease.

1.7.1 Targeted gene resequencing

Targeted or candidate gene resequencing studies are useful when there is clear evidence of variants presiding close to or in genes and these risk variants are more prevalent in persons with disease than in the overall population. Furthermore, the combination of variants observed in previous studies, such as GWAS, may not explain the entire genetic variance contributing to disease therefore further analysis into key risk genes can provide a more comprehensive architecture of genetic structure. Concentrating on coding SNPs has excellent potential for candidate gene analysis, as the number of coding SNPs is several magnitudes smaller than the overall number of SNPs in the genome (Cargill, Altshuler et al. 1999). This cost-effective approach maximizes the number of samples one can sequence compared to whole genome sequencing (WGS) by focusing on a limited number of candidate genes but in turn increasing the statistical power of finding a causal risk variant. This said, the cost of multiplex amplicon sequencing will effect the prioritization of

candidate genes, but successful strategies can be developed, such as ranking top those genes with the highest number of mutations (Fieuw, De Wilde et al. 2012). As well as saving on cost, the sequencing efficiency and high depth of coverage achieved (owing to sequencing the same gene or amplicon in many samples) allows much deeper examination of candidate genes for detection of both common and rare variants. Positive evidence for gene resequencing has implicated rare variation contributing to protective and causal phenotypic effects, for example individuals with low plasma levels of HDL cholesterol have been shown to inherit rare alleles in three candidate genes contributing to the Mendelian form of disease (Cohen, Kiss et al. 2004), whereas rare variants in the *IFIH1* gene were found to confer protection against T1D by altering protein expression and structure (Nejentsev, Walker et al. 2009). Similarly, Momozawa et al. identified low frequency coding variants from 63 GWAS-identified positional candidate genes showing protection against InBD in *IL32R,* but no rare variants were found predisposing to Crohn's disease (Momozawa, Mni et al. 2011). These studies, along with plenty others showing rare variation burden in disease provide compelling evidence of specific gene pathways predisposing to genetic disease (Johansen, Wang et al. 2010; O'Roak, Vives et al. 2012).

This research project involves a targeted resequencing component and this section is expanded on in the introduction of 'Chapter Five: Exome study candidate gene resequencing in 2,304 cases and 2,304 controls'.

1.7.2 Exome sequencing

The last five years has produced many successful genetic studies applying exome sequencing to explore coding variants that are not detected by GWAS arrays. Success in candidate gene and Mendelian disease studies have allowed the development of exome arrays for an entire search of the protein-coding region (~30 Mb) of the genome where a candidate gene is not known, making it more cost effective than WGS as only 1% of the genome is captured. As noted earlier, the concept surrounding exome sequencing is founded in population genetics theory which states that there is selection driving against amino acid

replacements in the protein-coding regions of the genome and so are rare in the population (Williamson, Hernandez et al. 2005; Yampolsky, Kondrashov et al. 2005; Kryukov, Pennacchio et al. 2007). Every gene is expected to contain functionally important variants that can be found and tested through exome sequencing, even if they are relatively rare because mutations are continually occurring in each protein-coding gene (for nonsynonymous variants the mutation rate per gene per generation is ~1 x 10^{-5}) (Nachman and Crowell 2000).

The interest in exome sequencing comes from the potential to identify many genes underlying complex traits and straightforward functional annotation of the coding variation. The method has proven extremely powerful for rare monogenic disorders (Byun, Abhyankar et al. 2010) giving much insight into disease etiology, for example in rare congenital diseases (Lin, Chen et al. 2012), and has progressed into the development of programs dedicated to the diagnosis of rare disease types (Maxmen 2011), coupled by ongoing developments in software for exome sequencing-based discovery (Li, Gui et al. 2012).

The first study to use exome sequencing was carried out by Ng et al. and highlighted the high number of coding variants in an individual exome: 12,500 variants of which 15-20% are rare in the population (Ng, Levy et al. 2008). A proof of principle study by the same authors captured 12 human exomes and developed a strategy with data from a small number of individuals to find a candidate gene for Freeman-Sheldon syndrome (Ng, Turner et al. 2009). A year later, exome sequencing was applied to ten unrelated probands with transmitting Kabuki syndrome and only found a suggestive candidate gene, *MLL2*, after less stringent filtering suggesting that diseases which are hetergenous may not necessary harbour the same mutations in the same genes across individuals or that not all the targeted exome was captured and sequenced (Ng, Bigham et al. 2010). The former point is a major caveat for complex disease.

Combining linkage with whole exome sequencing has become common practice; individuals contributing to a known linkage peak can either be exome

sequenced to find the causal variant, like in the analysis of *ADIPOQ* (Bowden, An et al. 2010), or exome variants can be filtered based on the linkage logarithm of odds (LOD) score from SNP based genome-wide linkage analysis and then on allele frequency depending on the inheritance model (Yamaguchi, Hosomichi et al. 2011). In terms of applying this method to a complex trait, evidence has shown that significent associations for complex disease, such as Crohn's and T1D, reside in the exons (Lehne, Lewis et al. 2011; Albrechtsen, Grarup et al. 2013). A myriad of success stories have been observed in autism spectrum disorders which have located de novo mutations subsiding in neurological gene pathways (O'Roak, Deriziotis et al. 2011; Sanders, Murtha et al. 2012).

Another main goal for overall understanding of genetic disease is to construct a mutation profile to deduce the mechanisms underlying disease progression, and such profiles can be started from birth (Christodoulou, Wiskin et al. 2012). In spite of this important issues must be considered when carrying out an exome sequencing study on a complex trait, such as the choice of samples to sequence, generation of sequence data to achieve a decent coverage over all protein-coding variants, and variant calling strategies (Do, Kathiresan et al. 2012). Data from the 1000G project phase one identified that the power to detect alleles at frequencies of >1% is equivalent between exome and low coverage sequencing but for rare alleles exome sequencing has much more power (Abecasis, Auton et al. 2012), so this is a suitable method to search for rare disease-causing mutations.

One major component of this research project is exome sequencing so this section is expanded in the introduction of 'Chapter 3: Exome Sequencing in 75 coeliac disease individuals'.

1.7.3 Whole genome sequencing

To truly elucidate the relationship between disease phenotype and their corresponding genetic basis, attention ought to be focused on the entire genome, where the entire spectrum of variants can be analysed. Many methods have been postulated on the best way to identify SNVs, insertion-deletions

(indels) and copy number variants (CNVs) in whole genomes sequences, but before variant calling, short reads can be assembled in two different ways: de novo assembly or reference panel assembly (Ng and Kirkness 2010). Once bionformatic methods are refined, WGS can be a powerful tool in disease research. Lupski et al. (2010) showed the power to diagnose a Mendelian disease using whole genome sequencing is much stronger than using targeted approaches (Lupski, Reid et al. 2010). WGS applications in inherited disease in families have narrowed candidate genes down to four in two Mendelian disorders (Roach, Glusman et al. 2010). This family based approach is low cost and the results can be extrapolated during investigation in other families with the same diseases. Other family based designs have shown de novo mutation hotspots that underlie autism (Michaelson, Shi et al. 2012). Furthermore low coverage WGS sequencing has been shown to give similar *P* values at known associated variants from genotype data (Pasaniuc, Rohland et al. 2012) indicating that low coverage WGS can yield several times the effective sample size of SNP array data with an increase in statistical power. A combination of GWAS imputation and WGS established robust disease associations in an isolated Icelandic population for sick sinus syndrome (Holm, Gudbjartsson et al. 2011). This study found a signal on chromosome 14 through GWAS and imputation of 1000G variants and then aimed to refine that signal by imputing 11 million WGS variants from seven cases and 80 controls into the GWAS dataset. The study showed that low population frequency (0.1-1%) variants can be found through this sequence-based association method and is a powerful approach for complex traits as only a subset of cases with well defined phenotypes were sequenced making it cost-effective (Zeggini 2011). With ongoing decrease in sequencing cost, WGS alone may be affordable in a large cohort of patients required for complex disease genetics. WGS has recently had a major push into clinical practice, especially in the UK. Government funding has propelled a £100 million cash injection into sequencing up to 100,000 patients with cancers and rare diseases. This massive sequencing effort will integrate cutting edge medical science into the current healthcare system and has the potential to create better-tailored treatments and therapy.

1.8 Summary and outline of research hypothesis and aims

This introduction chapter has summarized CD symptoms, epidemiology, physiology and immune characteristics, incorporating the role of how genetics had aided in finding up to 39 non-HLA disease associations. Common and low frequency (5-0.5%) variation in CD has been found through case control association studies, but there is a tail of weaker associations that are likely to explain some but not all disease heritability. Rare variant contribution of large effect is hypothesized to have an impact on disease onset. From the results of Mendelian studies and other complex traits, it is hypothesized that rare variation of large effect size contribute to the missing heritability of CD. These variants can be found through a combination of exome sequencing, rare variant genotyping and targeted gene resequencing methods. The next chapters will discuss exome sequencing in 75 CD individuals (Chapter 3), association based analysis with Illumina Immunobeadchip genotyping with protein coding SNPs generated from exome sequencing and linkage analysis in multigenerational multiply affected coeliac families (Chapter 4). Finally, from a combination of the outlined methods above, the resulting candidate genes have been taken forward in a targeted resequencing study of 2,304 cases and 2,304 controls (Chapter 5).

Chapter 2

General Methods

2.1 DNA sample collection

The DNA sample collection for the exome sequencing study (Chapter 3), entire Immunochip case control study and linkage analysis (Chapter 4) and the candidate gene resequencing study (Chapter 5) is described here.

An advertisement was placed in Coeliac UK magazine (England, Wales and Scotland registered charity) for coeliac sample collection. Upon receipt of interest, an Oragene© DNA saliva kit (with a unique identification number) supplied by DNA Genotek, Inc (Oragene© cat# OG-250) was delivered together with a consent form, questionnaire and research information sheet to each subject. Control DNA was obtained from relatives/friends of coeliac subjects without disease. Written informed consent was obtained from all subjects, with Ethics Committee/Institutional Review Board approval. All individuals are of European ancestry. Samples from ten multigenerational families used for exome sequencing and linkage analysis were collected from Paul Ciclitiria's coeliac family sample collection at St Thomas' Hospital, five families were from Susan Neuhausen at Beckman Research Institute at the City of Hope, California and one family was from Åsa Naluai at Gothenburg University, Sweden. In addition, DNA from blood samples were previously collected and extracted prior to my joining the research laboratory and this repository was available for use in this thesis project.

2.2 Genomic DNA extraction

For saliva DNA extraction, a protocol provided by DNA Genotek, (www.dnagenotek.com/DNA_Genotek_Product_Oragene_DNA_A_Lit.html) was followed. Briefly, Oragene© saliva pots were defrosted at room temperature for at least one hour and then placed in a 50°C incubator for a minimum of two hours. After incubation, 500μl of saliva was transferred into a labeled 0.5ml safe lock test tube (Eppendorf cat# 0030.121.023). The remainder was transferred into a 5ml cyrotube (Sigma Aldrich cat# CLS430663-500EA) and frozen at -80°C

for future use. 20μl of DNA purifier (Oragene OG-L2P) was added to 500μl of saliva, vortexed and incubated on ice for 10 minutes. The entire mixture was centrifuged at 13,000 rpm (15,000g) (Eppendorf Centrifuge 5415-D) for five minutes. The clear supernatant was transferred to 500μl of 100% ethanol and the remaining white pellet discarded. The tube was inverted ten times to mix, allowed to stand for ten minutes to allow DNA precipitation, followed by centrifugation at 13,000 rpm (15,000g) for two minutes. A visible DNA pellet was then washed with 200μl 70% ethanol after removal of the supernatant. The pellet was left to air dry for five minutes, and then re-suspended in 20μl of molecular biology grade water (Sigma Aldrich cat# W4502) and placed in a rack on an orbital shaker overnight. After overnight re-suspension, the extracted DNA was frozen at -80°C.

2.3 Genomic DNA quantification

All saliva extracted DNA and DNA received by St Thomas' Hospital was quantified using the Quant-iT™ picogreen DNA assay kit (Invitrogen cat# P11496). Quantification was carried out on 40 samples at any one time, arranged in 5 columns of 8 samples on a 96-well plate. 100x TE buffer was diluted to 1x and 999μl transferred to columns A1-H10 of a deep 96-well plate (VWR cat# 736-0344). A 1μl aliquot of DNA was added to the wells containing 1x TE buffer in duplicate (A1 and A2 for sample 1, B1 and B2 for sample 2, etc). The deep well plate was sealed and left on an orbital shaker to mix. DNA standards were prepared with lambda DNA and 1x TE buffer at the following concentrations: 2000ng/ml, 500ng/ml, 125ng/ml, 3125ng/ml, 7.81ng/ml, 1.95ng/ml, 0.488ng/ml, 0ng/ml (no lambda DNA). 100μl of DNA was transferred from the deep well plate into a 96-well fluorescence plate (Sigma Aldrich cat# CLS3610-48EA) and 100μl of each lambda DNA standard transferred in duplicate, starting from blank 1 x TE buffer. The supplied fluorescent dye was thawed and 50μl diluted in 11 ml of 1x TE buffer. 100μl of dye was transferred

to each well in the 96-well fluorescence plate, covered with foil and left for four minutes.

Fluorescence was measured on a fluorometer (POLARstar OPTIMA, BMG LabTech) with excitation filter set at 485P and emission filter set at 520P, and measurement set at 'FI top'. Results were collected on an excel spreadsheet; a standard curve was drawn with ng/ml of standard DNA on the y-axis and average fluorescence on x-axis. The best possible fit was ensured by only plotting values covering the range of fluorescence of the DNA samples being quantified. A polynomial trend line was fitted so r=1 or very close to 1, and a line equation was displayed and used to back-fit the standard curve fluorescenceas a quality check. The same equation was used to calculate the DNA concentration in each well; the average of duplicate samples was taken.

2.4 PCR and gel electrophoresis

Reagents, concentrations and volumes for a standard PCR experiment on human genomic DNA are outlined in table 2.1. Primers were received at 100mM concentration in water. Each primer pair was diluted to 10mM; 50μl of the forward stock primer and 50μl of the reverse stock primer was added to 400μl of molecular grade deionised water in an eppendorf tube. Stock human genomic DNA was diluted to 5ng/μl in molecular grade water and 5μl of this dilution was added to a PCR plate well. The PCR mix was prepared for *x* number of DNA samples, briefly vortexed, and 13μl added to each well containing DNA, plus a negative control. If PCR reactions with many different primers pairs were being performed, 2μl of each 10mM primer pair was added to the associated DNA sample. If not, the primer pair was added directly into the PCR mix. The final volume of each PCR reaction was 20μl and run at the following cycling times on a thermocycler: 96°C for 10 min, 35 cycles of 95°C for 15 sec, 72°C for 15 sec, 60°C for 15 sec, extension at 72°C for 5 min and final hold at 4°C for 10 min.

PCR products were viewed on a 2% agarose gel: 3g of agarose (Sigma Aldrich cat # A9539) was diluted in 150ml of 100x TAE buffer (Sigma Aldrich cat # T6025)

and heated in the microwave until fully dissolved (approximately 3 minutes). 10μl of GelRed™ nucleic acid stain gel (VWR cat # 89139-140) was added to the heated agarose, and the entire solution was poured into an electrophoresis tank and left to cool for 10 minutes. DNA samples mixed with 5μl loading dye were loaded onto the set gel along with 5μl of HyperLadder™ 100bp (Bioline cat # 33056); the gel was run at 100V for up to 20 minutes.

Table 2.1: Standard PCR reagents, concentrations and volumes for genomic DNA

Reagents (Supplier)	Concentration for one reaction	Volume for one reaction
Primers (Sigma Aldrich, 100μM)	10uM	2μl forward and reverse
AmpliTaq Gold® PCR Master Mix includes: AmpliTaq Gold® DNA polymerase Gold Buffer $MgCl_2$ (Life Sciences 1000 units cat # 4338858)	 5U/μl 10 x 25mM	 0.2μl 2μl 2μl
dNTPS (Bioline 100mM cat # 39025)	25mM	2μl
Stock genomic DNA	25ng	5μl
Molecular Grade Water (Sigma Aldrich cat # W4502)	-	6.8μl

2.5 Exome target capture

Two target capture methods are outlined in this section: the first method describes array capture with 12-sample pooling and multiplex sequencing (Phase One, intended as a pilot study); the second describes single sample capture with exome probes in-solution (Phase Two). The same library preparation reagents were used for both methods.

2.5.1 NimbleGen human exome 2.1M array (Phase one)

A pilot experiment was carried out with 60 samples on NimbleGen Sequence Capture Human Exome 2.1M Array, v1 (NimbleGen, USA), containing 180,000 coding exons (CCDS transcripts) and 551 miRNA exons totaling 26.7Mb of the human exome. Samples were multiplexed (12 samples per pool) and each sample in the pool had a unique 6bp barcode. A PCR was performed to incorporate the barcodes before capture to the array.

2.5.1.2 Library preparation and index PCR

Genomic DNA was prepared for high throughput sequencing by following the Illumina 'Preparing Samples for Multiplexed Paired-End Sequencing' Protocol (part #1005361 Revision B December 2008 available at www.illumina.com). Multiplex reagents were supplied by Illumina as part of the Multiplex Sample Preparation Oligonucleotide Kit (cat# PE-400-1001).

The Illumina 1005631 protocol outlines nebulization as a method for DNA fragmentation, however this application was modified to use sonication instead. 5μg of genomic DNA was diluted in a total volume of 50μl 1x TE buffer and 150μl of nebulization buffer. The BioRupter sonicator apparatus was assembled and filled with water and ice. Samples were loaded (6 per load) and sonicated for 30 minutes. Fragmented DNA was cleaned with QIAquick PCR purification kit (Qiagen cat # 28106) and 1μl was run on DNA 7500 bioanalyzer chip for size confirmation. The library preparation steps following DNA fragmentation were a) repairing the ends of double stranded DNA fragments; b) adding an A base overhang to the 3' ends of DNA fragments; c) ligation of DNA to double stranded DNA adapters. A purification step was performed with the QIAquick kit subsequent to steps a, b and c. Ligated DNA templates were size selected (between 250-300bp) and then a 6bp index (or barcode) was added via an 18 cycle PCR. After quantification on a spectrophotometer (Nanodrop, Nanodrop Technologies, USA), 12 indexed libraries were pooled together with the same concentration of each library in the pool.

2.5.1.3 Array capture and PCR

Array capture and post capture PCR was performed according to manufacturer's instructions (Nimblegen Sequence Capture User Guide: Sequence Capture Array Deliver v3.1, available at www.nimblegen.com)

Prior to microarray hybridization, a solution of Cot-1 DNA and 1ug of the pooled DNA library was dried down, resuspended in molecular grade water and then incubated at 70°C. Hybridization enhancing oligonucleotides (forward and reverse primers used in post capture PCR, added in excess, to prevent duplex formation between adapter regions of the genomic fragments) and hybridization buffers were added to the re-suspended pooled DNA and then denatured at 95°C. Samples were loaded onto a 2.1M sequence capture exome array and hybridized at 42°C for 66-72 hours on the Nimblegen hybridization system.

After hybridization, washing and elution of the post-captured library was performed; elution at 95°C with water was later replaced with elution with sodium hydroxide at room temperature (Nimblegen Sequence Capture User Guide: Sequence Capture Array Deliver v3.2, available at www.nimblegen.com); both methods were used in this experiment. A final PCR step on the eluate was carried out, where a reaction mix was prepared for 10 reactions per capture. For clean up after PCR, 5 reactions were pooled and added to 1250μl of PBI buffer (QIAquick PCR purification kit), and eluted in a total of 100μl EB buffer.

2.5.1.4 Enrichment qPCR

Enrichment qPCR was performed according to manufacturer's instructions (Nimblegen Sequence Capture User Guide: Sequence Capture Array Deliver v3.1, available at www.nimblegen.com). This assay determined whether capture was successful and used a standardized set of qPCR SYBR Green assays as internal quality controls for Nimblegen sequence capture experiments (referred to as NSC assays). The genomic loci recognized by the assays were included as capture targets on every human NimbleGen array and comparison by qPCR of

the relative DNA concentrations of the control loci in non-captured and captured samples allowed estimation of enrichment of a capture target. There were four NSC assays (NSC-0237, NSC-0247, NSC-0268, NSC-0207, refer to manufacturer's guide for primer sequences) and primers for each assay were diluted to 2um. Each non-capture and capture (post PCR) product was diluted to 5ng/μl in PCR grade water to use as qPCR templates. The following volumes of each component were added together to make a qPCR mix for one sample: 5.9μl PCR-grade water, 0.3μl NSC 2μm forward primer, 0.3μl NSC 2μm reverse primer, 7.5μl SYBR Green Master (2X), 1μl of 5ng/μl template (non-captured product, captured product or positive control genomic DNA templates). The following program was created on the Corbett Rotor-gene RT-PCR machine: pre-incubation 1 cycle 95°C 5 min; amplification 40 cycles 95°C 10 sec; melting curve 1 cycle of 60°C 1 min, 95°C 10 sec and 65°C 1 min; cooling cycle 40°C 10 sec. CT values for all reactions were collected and calculated for replicate reactions. For each different sample and NSC assay, the average CT value of captured template was subtracted from the non-captured template. This value was defined as the delta-CT. The fold enrichment was calculated for the NSC control locus by raising the PCR efficiency (or *E*: 1.84 for NSC-0237, 1.8 for NSC-0247, 1.78 for NSC-0268, 1.93 for NSC-0207) for that assay to the power of delta-CT measured for the corresponding control locus. This enrichment analysis was performed for every pool.

2.5.1.5 Pooled DNA library quantification

Prior to clustering and sequencing a sample on the high-throughput sequencing machine, the library was quantified by Nanodrop and assayed for size by running 1ul on Agilent 2100 Bioanalyzer with the DNA 7500 chip. The size of the DNA peak was determined by manual inspection of the resulting electropherogram. The molar concentration of the DNA was calculated using the formula below and the library diluted to 10nm for clustering:

$$y = \left[c \text{ x } (10^3) \text{ x } \left[\frac{1}{x \text{ x } 2} \right] \text{ x } \left[\frac{1}{s} \right] \right] \text{ x } 1000$$

where *y* = the molar concentration of the library (nM); *c* = concentration of DNA library in ng/µl; *x* = average molecular weight of a DNA base at 324.5; *s* = the resultant size of the DNA library. Samples were diluted to 10nM using the formula described.

2.5.2 Nimblegen EZ SeqCap human exome in-solution (Phase two)

A total of 75 single samples from coeliac individuals (DNA extracted from either blood or saliva) were prepared for high throughput sequencing and captured with 2.1 million Nimblegen EZ SeqCap human exome in-solution probes**,** covering 26.7Mb and 44.1Mb of the exome for the v1.0 kit and v2.0 kit respectively. Modifications to the protocol include the removal of pre capture PCR and replacing post capture LM-PCR with real time qPCR using SYBR green fluorescent dye. This required monitoring cycle-to-cycle amplification and executing PCR before it reached the amplification plateau. This modified protocol is in Appendix I-B where instructions for library preparation, in-solution captures and post capture PCR are outlined.

2.5.2.1 Single library quantification with qPCR

Single in-solution exome capture libraries were quantified with TaqMan qPCR, developed by Barts and the London Genome Centre. Previous libraries that generated optimum cluster numbers were used as controls. Target cluster numbers per sample were between 350-400K per tile with version 3 sequencing chemistry on the Illumina GAII*x*. Serial dilutions of 1ul per library and chosen controls were prepared with molecular grade water to give the following concentrations: 20x, 200x, 2000x, 20000x and 200000x. 5.5µl of TaqMan Universal PCR master mix (Applied Biosystems cat #4304437) and TaqMan

Tamra Probes (FAM reporter dye) (Applied Biosystems cat #450025) and 5ul of each triplicate of 2000x, 20000x and 200000x dilutions, plus a water control, were loaded onto a 384 well plate on an automated Biomex FX pre PCR robot, to give a final reaction mix of 10.5μl. The prepared 384 well plate was loaded onto the ABI 7900 HT Fast Real Time PCR system and run at the following cycling times: 50°C for 2 min, 95°C for 10 min, 60 cycles of 95°C for 15 min ad 60°C for 1 min. SDS 2.3 software was used to analyse results at absolute quantification by a ΔCT method.

2.6 Illumina Immunobeadchip genotyping

The entire exome sequencing cohort of 75 samples from phase two of the exome sequencing experiment plus 7728 UK coeliac cases and 8274 UK controls were genotyped on the Immunochip custom chip designed by Illumina. DNA was diluted to 200ng total with molecular grade water in a 96-well plate (Thermo Scientific cat #AB-0564). Genotyping was performed following the Illumina Infinium HD Ultra User Guide 11328087 Revision B (available at www.illumina.com) at Barts and the London Genome Centre. Genomic DNA was whole genome amplified without PCR, by overnight incubation. After fragmention, precipitation and resuspension, each DNA sample was hybridized to the custom beadchip in a capillary flow-through chamber. Non-specific DNA was washed away and then stained for single base extension of the oligonucleotides present on the beadchip. As the captured DNA is used as a template, the detected label is incorporated onto the beadchip, and thereby determining the genotype of the sample. All beadchips were analysed on the Illumina iScan housed at the Institute of Health, University College London.

2.7 Fluidigm 48.48 Access Array Integrated Fluidic Circuit technology

Fluidigm Access Array™ Integrated Fluidic Circuit (IFC) technology was used to for target specific amplification of 26 exome sequencing candidate genes in 2,304 cases and 2,304 controls, equating to three 1,536-multipex libraries. DNA

was collected, extracted and quantified as stated in sections 2.1-2.3 and diluted to 50ng/μl; 1μl of this dilution was transferred to a 96-well plate. Instructions for sample amplification and barcode PCR were followed according to the manufacturer's user guide (Access Array System for Illumina Sequencing Platform, P/N 100-3770, Revision C1).

There are three sets of Fluidigm machines: a Pre-PCR IFC Controller to load samples and primers into the IFC array, an FC1 cycler for PCR amplification and Post-PCR IFC Controller to harvest all PCR products. All three machines are housed at Barts and the London Genome Centre.

2.7.1 Assay design and pooling

The NCBI gene symbol for 24 target genes were sent to the Fluidigm assay design team who designed primers around each target exon (excluding 5'UTR and 3'UTR). The team was instructed to design amplicons with an optimal length of 200bp (minimum 150bp, maximum 200bp). 506 assays were designed and each assay was tagged with CS1/CS2 (common sequence) sequences for use with the 1,536 barcodes. Six separate 96-well plates containing the assays and excel sheets detailing the plate layouts were received with the Fluidigm ID number 1488AAP1201.

A stock multiplex primer mix was prepared according to manufacturer's instructions (Fluidigm Access Array Multiplex 20x Primer Solution Preparation Quick Reference, PN 100-3895, Revision B1). Briefly, 90μl from each of the 12 columns across plates 1-6 was transferred to a designated single column on the stock plate i.e. columns A1-H12 from primer plate 1 were pooled into column 1 in the stock plate, columns A1-H12 from primer plate 2 were pooled into column 2 in the stock plate, and so forth. All primer pairs were provided at 60μM and after combining all 12 wells for each plate into the stock plate, each primer was at a final concentration of 5μM. From this stock plate, 20μl from each column (A1 – H6) was transferred into a working stock plate, along with 5μl of 20x Access Array loading reagent and 75μl of 1x TE buffer. This was labeled as the 20X primer solution plate.

2.7.2 Multiplex PCR on the Access Array

DNA was prepared for multiplex PCR on the Fluidigm F1 cycler according to the manufacturer's instructions (Access Array System for Illumina Sequencing Platform, P/N 100-3770, Revision C1). In brief, an Access Array IFC was injected with control line fluid and 500μl of harvest solution was added into H1-H4 wells on the IFC. The array was loaded onto the Pre-PCR IFC Controller and primed for 10 minutes. A PCR mix combining components from the FastStart High Fidelity PCR system kit (Roche cat # 04738292001), 20X Access Array loading reagent and water was prepared for 96 samples (two IFC arrays was required for one 96-well DNA plate). 4μl of PCR mix was added each DNA sample (50ng) in the 96-well plate; the sample mix plate was vortexed and spun down for 30 seconds. The IFC was loaded with 4μl of the sample mix solution into each of the sample wells and 4μl of the 20x primer solution into each of the primer wells. The IFC array was loaded onto the Pre-PCR IFC controller and the 'Load Mix; script was run for 1 hour and 6 minutes. After the IFC was loaded, it was transferred to the FC1 cycler in the post PCR laboratory; the AA 48x48 Standard v1 protocol was selected for PCR amplification. After the PCR had completed, the IFC was harvested on the Post-PCR IFC Controller and PCR products were transferred into a labeled 96-well plate. Six samples from each harvest plate (3 samples per array) were checked on the Agilent 2100 Bioanalyzer with the DNA 1000 DNA chip.

2.7.3 Barcode PCR

A barcode PCR was performed to attach 1,536 barcodes to each sample and the protocol was followed according to manufacturer's instructions (Access Array System for Illumina Sequencing Platform, P/N 100-3770, Revision C1). Fluidigm supplied all barcodes and 4μl of each barcode was aliquoted into 16 separate 96-well plates previously. In brief, a PCR mix combining components from the FastStart High Fidelity PCR system kit was prepared for 96 samples. A 100-fold dilution of the harvested PCR products was prepared by aliquoting 1μl into 99μl

of water. 1μl of this dilution was added to 15μl of PCR mix and 4μl of barcode, giving a 20μl PCR reaction. The PCR plate was placed on a PCR thermal cycler and the following protocol was run: 95°C for 10 min, 15 cycles of 95°C for 15 sec, 60°C for 30 sec, 72°C for 1 min, final extension at 72°C for 3 min. Six samples from each barcode PCR plate (3 samples per array) were checked on the Agilent 2100 Bioanalyzer with the DNA 1000 chip.

2.7.4 Post PCR purification, quantification and library pooling

All 1,536-multiplexed barcode samples were pooled into one library containing 1μl of each sample. The entire library was purified following manufacturer's instructions (Access Array System for Illumina Sequencing Platform, P/N 100-3770, Revision C1). In brief, 12μl of sample pool was added to 24μl of TE buffer and 36μl of Ampure XP beads (Agencourt AMPure XP, Beckman Coulter cat # A63880) in an eppendorf tube. After a 10-minute incubation at room temperature, the tube was placed on a magnetic separator and beads were allowed to separate from the supernatant. Once the supernatant was clear, it was removed and discarded. The beads were washed with 180μl of 70% ethanol twice and then air dried for 10 minutes. 40μl of DNA suspension buffer was added to the beads and the tube was placed on the magnet. The supernatant was transferred to a new-labeled tube and 1μl was run on the Agilent 1000 DNA chip to check the size and concentration of the library. The library was diluted to 10nM and stored at -20°C.

2.8 High throughput sequencing on Illumina Genome Analyzer II*x*, MiSeq and HiSeq 2000

All pooled and single sample exome libraries were sequenced on the Genome Analyzer II*x* at Barts and the London Genome Centre. All 1,536-multiplex Fluidigm libraries were sequenced on the Illumina MiSeq at Barts and the London Genome Centre for quality control purposes and then on the HiSeq

2000 at the NIHR GSTFT/KCL Biomedical Research Centre at Guy's Hospital. The sections below outline cluster generation and paired end sequencing for the Illumina GAIIx, sample preparation for MiSeq sequencing and HiSeq 2000 run settings.

2.8.1 Cluster generation for Illumina GAII*x*

All cluster generation was performed on the Illumina Cluster Generation System using the paired-end flow-cell v4 and paired-end cluster generation kits for genomic DNA sequencing v4 (Illumina, USA). The appropriate amount of DNA was loaded onto each flow cell lane according to quantification results; normally 4pM is the required amount. A single control lane of bacteriophage ϕX-174 DNA (PhiX DNA, PhiX control kit, Illumina USA, CT-901-2001) was run on each flowcell.

2.8.2 Paired end and multiplex sequencing on Illumina GAII*x*

Subsequent to cluster generation, multiplex sequencing (for array capture, phase one) and single sample sequencing (for in-solution capture, phase two) was performed on the Illumina Genome Analyzer II*x*. The paired-end module was used for paired-end sequencing. A 76bp paired-end sequencing run was performed for phase one single samples using a combination of 36bp cycle sequencing kits v3 and v4. Phase two multiplex-pooled samples were also sequenced 76bp paired-end, with an additional 6bp index step.

2.8.3 Illumina MiSeq and HiSeq2000

A 50bp paired-end 11bp index MiSeq run was performed for each 1,536-multiplex library with version 1 and 2 sequencing kits. The sample library was prepared according to manufacturer's instructions (Illumina Sample Preparation for MiSeq, p/n 15028881 Revision A). In brief, the 10nM library was diluted in elution buffer to 2nM. A stock of sodium hydroxide was diluted to 0.1N and

10μl was added to 10μl of 2nM DNA. The mix was incubated for 5 minutes to allow the double stranded DNA templates to denature. 800μl of supplied HT1 buffer was added to the denatured DNA to make a 20pM solution. This was further diluted to 4pM. 600μl of 4pM DNA library was loaded into the sample well on the reagent cartridge.

CS1 and CS2 custom LNA™ sequencing primers and CS1rc and CS2rc index custom LNA™ sequencing primers were supplied by Exiqon (www.exiqon.com) and resuspended in low EDTA TE buffer (10 mM Tris pH 8. 0.1mM EDTA) for a final concentration of 100μM each. CS1/CS2 and CS1rc/CS2rc were combined to make stock FL1 and FL2 primers, with a final concentration of 50μM for each primer. The primers were spiked into the MiSeq reagent cartridge: 7 μl of FL1 into well 12 for read one, 7 μl of FL2 into well 13 for the index read and 7 μl of FL1 into well 14 for read two (see Table 2.2 for primer sequences). The run was started after all software checks were complete.

All three 1,536-multiplex libraries were run on one HiSeq flow cell (one library per lane) using a TruSeq v3 sequencing kit (Illumina cat # FC-401-3002). The run length was 101bp paired-end with an 11bp index read. Primers FL1 and FL2 were also spiked into the libraries for the forward and index reads.

Table 2.2: Fluidigm oligonucleotide sequences for Illumina MiSeq and HiSeq 2000 sequencing

Primer Name	Oligonucleotide Name	Sequence (5' -3')
FL1	CS1	A+CA+CTG+ACGACA TGGTTCT ACA
	CS2	T+AC+GGT+AGCAGAGACTTGGTCT
FL2	CS1rc	T+GT+AG+AACCATGTCGTCAGTGT
	CS2rc	A+GAC+CA+AGTCTCTGCTACCGTA

LNA nucleotides preceded by a "+"

2.9 DNA sequence alignment and variant annotation

2.9.1 Phase one and two exome sequencing study

A custom script was used to trim primer sequences from 150 fastq files (read 1 and read 2 data for 75 samples). The trimmed files were aligned to hg18/build37 of an indexed human genome using the short read mapper, Novoalign with settings "-H -k -a -o Soft -t 250". Novoalign allows lane specific alignment-based base call quality calibration and uses base call qualities in alignments, and permits alignment of small indels (15bp) in both reads. The Needleman-Wunsch algorithm was used for paired end data. Samtools v0.1.16, VCFTools v0.1.5 and PicardTools v1.55 were used for data processing. SNP and indel annotation was performed with both SeattleSeq and Annovar annotation software.

2.9.2 Fluidigm pilot and candidate gene resequencing study

All PCR amplicon sequencing oligonucleotides were trimmed from 9,216 fastq files (read 1 and read 2 data for 4608 samples) using a modified version of Btrim software (Kong 2011) with the following settings: "-k -S -u 2 -a -100". Trimmed fastq files were aligned to hg19/build37 of an indexed human genome using Novoalign with the following settings: "-t 100 -H -F ILM1.8 -g 65 -x 7 -o FullNW -c 1". Sorted and indexed bam files were created using Samtools v0.1.18. SNP and indel annotation was performed with the genome analysis toolkit (GATK) v2.3-9. Firstly, interval files containing all the exon coordinates for the target specific amplicons were generated and used for realignment around known indels (1000 genomes and Mills-Devine 2-hit) and sample level novel indels. This step was done to eliminate potential false positive SNPs caused by hidden indels. The base quality scores of all realigned files were then recalibrated to reduce bias.

SNPs and indels were called with the following GATK settings using the 'Unified Genotyper' option:

For SNPs:

'--min_base_quality_score 15 -stand_call_conf 30 --baq CALCULATE_AS_NECESSARY -glm SNP --baqGapOpenPenalty 65 --downsampling_type BY_SAMPLE --downsample_to_coverage 250' and then hard filtered using GATK settings 'QUAL<80.0 DP<20 MQ<40.0 QD<2.0 MQRankSum<-12.5 HRun>5'

For indels:

'--min_base_quality_score 15 -stand_call_conf 30 --baq CALCULATE_AS_NECESSARY -glm INDEL --baqGapOpenPenalty 65 --downsampling_type BY_SAMPLE --downsample_to_coverage 250' and then hard filtered using GATK settings 'QUAL<80.0 DP<20 QD<2.0'

Depth of coverage analysis was performed using GATK v2.3-9. Annotation of all variants was performed using the GENCODE V14 dataset, an ENCODE sub-project with annotated variants from all protein coding loci, non-coding loci (with alternatively transcribed variants and transcripts) and pseudogenes (Howald, Tanzer et al. 2012). Coding and functional variants were identified. All statistical analysis, including variant summary statistics and rare variant association and gene burden tests was performed using PLINK/SEQ v0.09.

Chapter 3

Exome sequencing in 75 coeliac disease individuals

3.1 Introduction

This chapter details a study of exome sequencing in 75 coeliac individuals from large and small CD pedigrees. So far, 39 non-HLA loci associated with CD risk have been located through GWAS, but collectively only explain 13.7% of disease heritability (Trynka, Hunt et al. 2011). This is because, other than the CDCV model, standard GWAS are not powered or designed to detect all variation based on other models attributed to disease risk across the allele frequency spectrum: infinitesimal model (common variants of relatively small effect), rare allele model (rare variants of large effect), the broad sense heritability model (combination of environment and epigenetic interactions) (Gibson 2011). The largest meta-GWA studies in body mass index and height have shown it is unlikely that more than a few hundred loci containing variants with allele frequencies of >5% will ever be confirmed for most diseases, and these may not explain even half of the genetic variance (Lango Allen, Estrada et al. 2010; Speliotes, Willer et al. 2010). As there are several causal variants on a common haplotype, of which can be in imperfect LD with the genotyped markers, GWAS SNP markers still underestimate disease-associated risk (Manolio, Collins et al. 2009). Hence, moving away from GWAS-based designs and targeting variants of lower frequency are required to locate the remaining genetic variance required to calculate missing disease heritability, and for this exome sequencing can be an effective method.

Exome target capture coupled with NGS are able find unknown functional genetic variants that occur infrequently in the population or are private to the affected individual. Firstly, it is largely acknowledged that coding mutations are indeed present at low frequencies in the population; mutations in these regions effect the expression of proteins, sustaining a deleterious effect on their function, and so are prevented from attaining a high frequency by selection (Fay, Wyckoff et al. 2001; Bustamante, Fledel-Alon et al. 2005). Present discoveries have substantiated this by highlighting that rare SNVs are more likely to be functional than common ones, and these rare functional SNVs are found at lower population allele frequencies (Zhu, Ge et al. 2011). Furthermore,

slightly deleterious alleles have been found to be younger, higher in abundance and population-specific compared to neutral alleles at the same frequency (Kiezun, Pulit et al. 2013). Secondly, exome sequencing is better designed to test the rare (typically MAF <0.5%) variant-common (typically MAF >5%) disease model, which states that many rare alleles of large effect are largely responsible for disease; this hypothesis has been proven in complex quantitative traits showing evidence of involvement of a few rare (1-5% allele frequency) and many ultra-rare/near-private mutations in disease genes (Cohen, Kiss et al. 2004; Romeo, Pennacchio et al. 2007; Ji, Foo et al. 2008). Finally, the effects of deleted exons and premature stop codons on protein function can be easily translated, not only in rare disease but also for complex common disease without a clear mode of inheritance. For example, rare (MAF <3%) protective *IFIH1* mutations against T1D suggest the causative factor to be an enterovirus (Nejentsev, Walker et al. 2009), whereas autophagy has been recognized as a new mechanism for Crohn's disease resulting from mutations in *NOD2/CARD15* and *ATG16L1* (Homer, Richmond et al. 2010).

These advantages offer great potential for complex studies but only if performed on a sample set powerful enough to detect rare variants. For quantitative traits, selecting samples from extreme ends of trait distribution can help enrich for disease causing variants, but as CD is a binary trait exhibiting a dichotomous phenotypic expression, it is difficult to stratify extreme cases and controls. In this case, making use of multiply affected families can offer an opportunity for the pursuit of novel variant discovery since extreme familial clustering might imply disease risk variants of higher penetrance (Bodmer and Bonilla 2008). Recent suggestions have focused on family-based analyses for increasing power, as truly private mutations are more likely to be pathogenic and more likely to cluster in families with disease (Kazma and Bailey 2011; Do, Kathiresan et al. 2012). A family-based design can potentially enrich the sample for very rare variants for which the effect would be concealed at the population level. This design can also filter hundreds of mutations to a few potential candidate mutations shared by affected related individuals.

Although there is no current example of a high risk rare mutation in CD, Crohn's disease which has similar heritability provides an example: three major causal *NOD2* mutations of population allele frequency ~1-3% and homozygous genotypes conferring odds ratios of ~15 for disease susceptibility under an additive model (Lesage, Zouali et al. 2002; Economou, Trikalinos et al. 2004). Other examples of rare coding variants in common disease are *TREX1* in systemic lupus erythematosus, *IFIHI* in T1D, *CARD14* in psoriasis and *ANGPTL4* in HDL cholesterol levels (Lee-Kirsch, Gong et al. 2007; Romeo, Pennacchio et al. 2007; Nejentsev, Walker et al. 2009; Jordan, Cao et al. 2012). Locating such rare variant(s) in a novel gene(s) or in one where previous common intronic GWAS-risk variants have been located, will allow insight into genetic variation in CD that is not yet accounted for. Additionally, focusing on immune genes is a logical strategy based on known coeliac and other overlapping autoimmune disease associations that have been discovered in immune mediated pathways.

To further understand where the missing genetic variation (and hence heritability) in CD lies, an exome target capture and high throughput sequencing experimental design with a family-based sample set was chosen for this study. Many target capture kits are commercially available for targeted resequencing large portions of the genome with NGS technology, proven to be a powerful approach for identifying genomic variation associated with disease (Ng, Turner et al. 2009; Ng, Bigham et al. 2010; Ng, Buckingham et al. 2010). The ultimate goal of resequencing is to accurately identify these variants in a cost-effective manner, whilst obtaining uniform and adequate sequencing read depth across the target region to sufficiently call variants. Roche NimbleGen™ was the first commercial company to release target capture reagents for 27.6Mb of the human exome using a microarray, used in phase one of this study. They later released a solution capture kit used for the majority of data generation in phase two of this study. Subsequently, Agilent is now the favoured capture choice but this kit was not available at the time of experimental work (April 2009).

3.2 Aims and hypothesis

The hypothesis for this study is that moderately highly penetrant rare (defined here as MAF <5%) variants of large effect size (allelic odds ratios ~ 2 - 5), possibly missed by family linkage studies using a small number of markers and/or families and common variant GWAS, predispose to CD risk and these variants account for a proportion of the missing heritability of disease.

A summary of the project aims are outlined here:

a) To identify, across the entire protein coding genome, rare coding variants that might directly affect gene/protein function in individuals with disease.

b) To target capture and sequence the exomes of distantly related affected individuals from large multiply affected families for the enrichment of private segregating disease causing mutations pointing toward a potential monogenic form of disease. Some multiply affected families will be sufficiently powered to perform segregation analysis to detect novel, potentially causal, variants.

c) To target capture and sequence the exomes of individuals with an extreme disease phenotype. An extreme phenotype individual can be described as having early disease onset, severe symptoms or belonging to a family with a high incidence disease rate. The latter has been chosen here in the hope that more disease causing mutations are prevalent in an individual from a coeliac family.

d) To perform shared variant-based analysis in related exomes and case-control rare variant analysis across the entire dataset (b and c combined) to increase the likelihood of finding more high-risk rare LoF variants.

e) To create a list of candidate genes for further targeted resequencing in a larger number of cases and controls.

3.3 Experimental design and sample selection

The experimental design is split into phase one, intended as a pilot study, and phase two where all data generated was used for downstream analysis and novel variant detection. Affected coeliac individuals were diagnosed according to standard clinical, serological and histopathological criteria, including small intestinal biopsy.

Phase one: multiplex exome sequencing with microarray capture

The phase one pilot study consisted of Roche NimbleGen™ exome microarray target capture and lllumina GAII*x* multiplex high throughput sequencing of 60 unrelated coeliac individuals with young age at disease onset (between 1 and 34 years of age). The main focus here was to assess data quality from a newly released commercial microarray target capture product by Roche NimbleGen™ based on: i) the number of usable reads per multiplex pool by determining the number of unique non-duplicate (or non-clonal) reads; ii) concordance rates with array-based genotyped SNPs; iii) whether multiplexing would effectively increase sample throughput without compromising read depth for each sample.

Phase two: single-sample exome sequencing with in-solution capture

The Roche NimbleGen™ in-solution target capture kit was used in phase two of the study on a sample set of 75 coeliac individuals followed by Illumina GAII*x* single-plex high throughput sequencing. The 75-case sample set included:

- 35 individuals from large (>2 generations) multiply affected coeliac families: two or more exomes were sequenced per family.
- 40 unrelated individuals from smaller affected coeliac families: one exome was sequenced per family.

All samples were from the UK, USA and Sweden (Appendix I-A details overall sample populations). The primary focus for phase two was to selectively enrich for disease causing mutations by sequencing relatives from large coeliac affected families, with the assumption that closely related affected individuals are more likely to share rare highly penetrant mutations. Additionally, sequencing individuals from smaller disease families will provide an enrichment of potentially causal rare variants in the dataset.

3.4 Phase One: multiplex exome sequencing with microarray capture

3.4.1 Phase One: Laboratory and in silico methods

DNA extraction, quantification, microarray exome target capture, sequencing, data alignments and variant annotation methods for phase one are specified in Chapter 2: General Methods. In brief, 5μg of genomic DNA from each sample was processed to create a DNA library, which consisted of DNA fragmentation, end polishing, adapter-ligation and pre-microarray capture index PCR. Twelve indexed samples were pooled together and then hybridized to exon probes on a microarray, consisting of 26.7Mb of the human exome (NimbleGen™ Sequence Capture Human Exome 2.1M Array version 1.0). A second PCR to enrich the eluted sample pool was performed and then a quantitative PCR (qPCR) was performed on the captured and non-captured samples to estimate relative fold-enrichment. All five pools (one flow cell lane per multiplex pool) were sequenced 76bp paired end, 6bp index, on the Illumina GAII*x* at Barts and the London Genome Centre. Samples were aligned to hg18/build36 of an indexed human genome using the short read mapper, Novoalign (www.novocraft.com). Variants were called with a custom Bayesian SNP caller. The general premise in detecting variants using a Bayesian method includes assigning quality (Phred) scores to mapped reads and then if there is a sufficient number of high-quality allele differences between the reference and sample genomes, the SNP is called. The Bayes' theorem is then applied,

$$P(E_i|D) = \frac{P(D|E_i)P(E_i)}{P(D)} = \frac{P(D|E_i)P(E_i)}{\Sigma_j P(D|E_j)P(E_j)}$$

where $P(E_i|D)$ is the posterior probability of event E_i given the observed data D, the prior probabilities $P(E_i)$ and the conditional probabilities $P(D|E_i)$ (You, Murillo et al. 2012). Using the Bayes theorem one can determine the SNP genotype with the highest posterior probability at each site.

At this stage, annotation was performed using SeattleSeq annotation software (www.gvs.gs.washington.edu/SeattleSeqAnnotation).

3.4.2 Phase One: Results

Data for the microarray capture experiment on five pools containing twelve individuals per pool was assessed for high quality sequencing reads. Table 3.1 shows the PCR cycling conditions pre and post-capture, the number of total and uniquely aligned reads, the number of clonal reads and the enrichment percentages per pool.

3.4.2.1 Indexing, enrichment and clonal reads assessment

It was important to assess whether there was even coverage of index (barcode) reads across each sample in the multiplex pool - a range of variable reads causes difficulty in downstream analysis especially when calling variants. Pool four was tested for how well the multiplexing worked. The range of perfectly tagged read-pairs passing filter was 89.3% - 95.7%, however the index counts for each sample were somewhat variable (Figure 3.1).

Successful enrichment of target exons from human genomic DNA was evaluated by four real-time qPCR control targets, ranging from 39.5% - 50% of captured reads mapping to within 500bp of the target exon (Table 3.1). Using pool four as an example, 2,215,692 of 4,431,385 reads map within 500bp of array target regions (50%). However, 17.9% of reads are clonal duplicates, and this number was higher for other pools (Figure 3.2). Clonal reads are multiple reads with the same orientation, start position and read length, and arise from PCR amplification. Although PCR amplification increases the number of available molecules for sequencing, random errors can be introduced due to changes in the number and representation of template molecules. For pool one, 13,791,191 out of a total of 19,942147 uniquely aligned reads were clonal resulting in only 30.8% usable reads. This pool had 18 cycles of pre-capture index PCR and 20 cycles post-capture PCR. In contrast, pool four had less cycles of post-capture PCR (15 cycles) and a much higher number of non-duplicates (82%). From these results, it was clear that less clonality correlated with a reduction in post-capture PCR cycles.

Figure 3.1: Bar plot of the number of reads per index

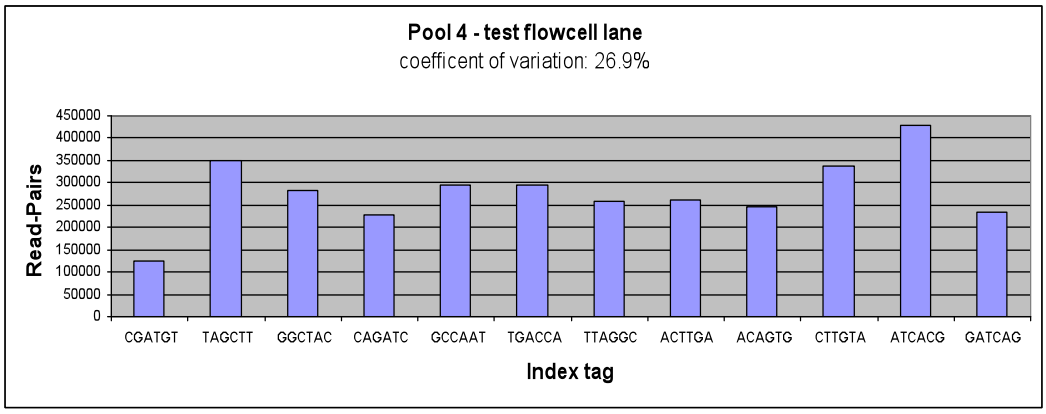
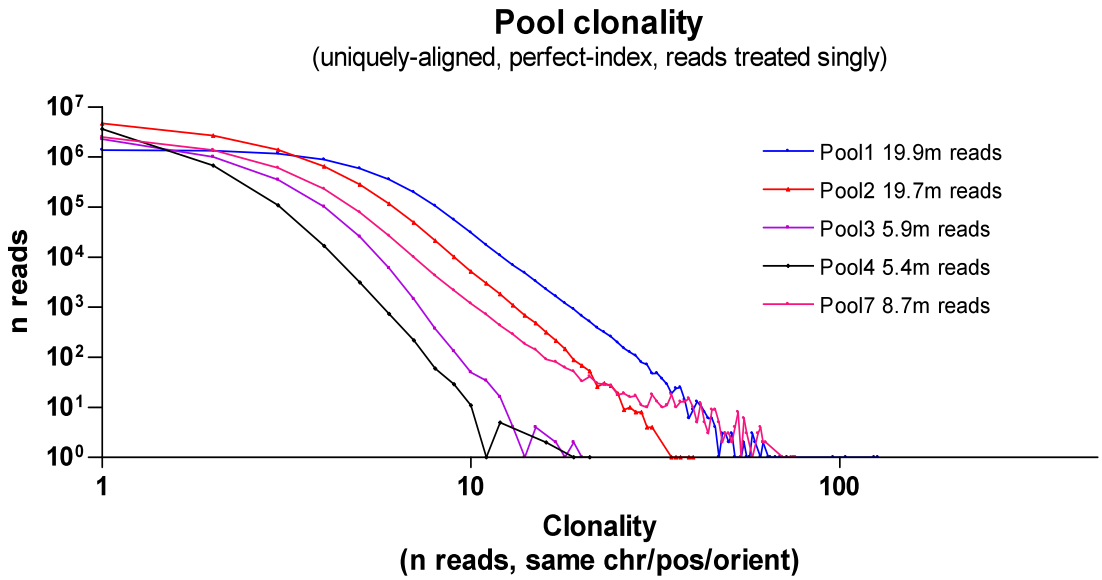


Table 3.1: PCR conditions, number of clonal reads and enrichment statistics for five multiplex pools

POOL	One	Two	Three	Four	Seven
PCR conditions and DNA input for array hybridization	18 cycle index PCR 1µg pool hyb 20 cycle post PCR	18 cycle index PCR 1µg pool hyb 20 cycle post PCR	18 cycle index PCR 1µg pool hyb 20 cycle post PCR	18 cycle index PCR 1µg pool hyb 15 cycle post PCR	18 cycle index PCR 3µg pool hyb 20 cycle post PCR
Total Reads	23,164,598	26,210,605	7,057,059	7,286,924	10,220,650
of which perfect 6bp index AND uniquely aligned	19,942,147	19,647,290	5,936,933	5,395,104	8,732,304
of which perfect 6bp index AND uniquely aligned AND non-clonal	6,150,956	9,898,078	3,777,536	4,431,385	3,885,875
% clonal reads	69.2%	49.7%	36.4%	17.9%	44.5%
% enrichment (map ± 500bp of target exon	46.3%	43.7%	45.0%	50.0%	39.5%

Figure 3.2: Graph of the number of clonal reads in each uniquely aligned pool



3.4.2.2 Sequence-based calls versus genotype-based calls concordance rate

Most sequencing data was obtained for sample CAP152639 from pool two (index CAGATC). For this sample, 1.6m reads were uniquely aligned and non-clonal, equating to 90Mb of total sequence. 21.4Mb of this was quality filtered sequence data at exome bases (base call quality higher than phred score of 10, a maximum of 5 mismatches per read and ~1 x mean read depth across the exome). SNP calls for sample CAP152639 were compared to array based genotype calls from Hap300 bead chips (containing 5,869 coding SNPs). 169 exome sequenced SNPs out of 172 array genotyped SNPs had identical calls, giving a 1.7% error rate at ≥ 4x read depth. There were two sampling errors that contributed to this error rate: one was no observation of the second allele at a heterozygous SNP and the second was a sequencing miscall error.

3.4.3 Phase One: Conclusions

This experiment was performed to assess data quality using a new commercial target capture kit developed by Roche NimbleGen™ for 27.6Mb of the human exome. The protocol outlined a two-step PCR scheme, one to introduce 6bp indexes and a second to amplify the eluted pooled product. It was evident that duplicate template molecules were prevalent in the sequencing data due to excess PCR; this observation has been previously highlighted (Kozarewa, Ning et al. 2009; Kozarewa and Turner 2011). Although the number of reads for variant calling was reduced, a high concordance rate with genotype-based calls indicated that the capture method worked well. One disadvantage of multiplexing was the variability of read-pairs per sample per pool, and a consequence of this was less specificity of variant calls across the dataset. Based on these results a number of changes were made to the methodology for phase two of the study:

1. *No multiplexing*: this is an attractive option to sequence multiple individuals in one run, reducing cost and time, but numbers of reads were not equimolar for all twelve individuals per pool and limited the number of sequence reads for certain samples, thus limited the number of SNP calls. Single-plex sequencing was chosen for phase two.
2. *Minimizing PCR cycles*: all NGS applications use PCR in library preparation methods, but the amount of PCR can be altered to reduce sequencing error and duplicate reads. The number of cycles used in the post-capture PCR step to enrich the eluted sample was minimized by qPCR in phase two. PCR allows exponential amplification of a DNA template, but once PCR components become limiting (such as DNA template) a plateau effect occurs in late PCR cycles meaning that amplification no longer proceeds in an exponential manner. This reduces amplification efficiency and can produce duplicate reads because the extra number of cycles amplifies the same sequence multiple times. qPCR uses fewer amplification cycles than a basic PCR because the measurement is taken prior to the reaction plateau.

3. *Improving DNA fragmentation*: the machine used to perform DNA fragmentation method was changed from the BioRupter to a Covaris LE220, a focused-ultrasonicator machine. The Covaris fragmentation process produces a much tighter range of DNA fragments, crucial for library preparation, in a faster processing time (200bp in 3 minutes and 300bp in 80 seconds).
4. *Increasing on target reads*: flow-cell cluster density was optimized by measuring DNA library concentration by qPCR instead of Nanodrop.
5. *Improve target capture*: an improved capture method was released towards the end of the phase one experiment, which promised better capture success and better automation. Roche Nimblegen™ SeqCap EZ Human Exome In-Solution capture eliminated the use of a hybridization station (used for array hybridization). Single in-solution captures were performed in a standard thermocycler allowing higher throughput. This method was also easily scalable using a multiple sample magnet for capturing the biotinylated DNA oligonucleotidess with streptavidin beads, unlike array captures where only one array could be processed at once.

3.5 Phase Two: single-sample exome sequencing with in-solution capture

3.5.1 Selecting related samples for exome sequencing

Relatives from the same family were selected for exome sequencing based on the amount of genetic sharing with the related individual and if the shared segments were likely to contain any disease risk mutations. Plink analysis on genotype data for two first cousins once removed (SAL-12757-0 and SAL-13281-5) was performed to identify shared chromosomal regions and confirmed that 6.25% (the expected number for this relationship type) of 19 long segments of DNA (366.588Mb) was shared. Further analysis highlighted that 99 out of 180 rare calls in the shared interval in SAL-12757-0 were also seen in SAL-13281-5, and 99 out of 199 rare calls in SAL-13281-5 were also seen in SAL-12757-0. This

supports the assumption that closely related affected individuals share rare mutations. The analysis also answered the question of whether having to sequence the second individual would be beneficial: at 6.25% sharing there were ~200 rare mutations (<5% MAF based on 1000G) in shared intervals (as defined by GWAS), and 100 of them were shared with the other individual in the pair, so exome sequencing the second individual is beneficial. Going less than 6.25% sharing i.e. sequencing second (3.125% sharing) and third cousins (0.781% sharing) decreases the size of shared intervals and the number of mutations to screen but a higher degree of sharing i.e. siblings, increases the chances of sharing high-risk variants but the number of shared mutations becomes too large. One individual in each affected pair with sharing in the 3-10% range was targeted to obtain a manageable number of mutations to assess for high disease risk. In conjunction with the increased likelihood of finding a novel risk mutation, this approach subsequently relaxes potential consequences of genetic heterogeneity by only considering novel variants in a given gene shared by related exomes as candidates.

3.5.2 Phase Two: Laboratory and in silico methods

DNA extraction, quantification, exome capture, sequencing, data alignments and variant annotation methods for phase two are specified in Chapter 2: General Methods. The full protocol for this method is outlined in Appendix I-B. In brief, the same library preparation process used in phase one was carried out on 75 single sample exome captures, with some protocol alterations. Samples were not multiplexed so the pre-capture index-PCR step (to add on a barcode) was removed. A standard post-capture PCR was replaced with real time qPCR using SYBR green; cycle-to-cycle amplification was monitored and stopped at 13-15 cycles (before it reached the amplification plateau). 71 single sample DNA libraries were hybridized to version 1 in-solution exon probes (27.6Mb human exome) and four samples were hybridized to version 2 (44.1Mb human exome) exon probes (Nimblegen SeqCap EZ Human Exome In-Solution). Each sample was sequenced 76bp paired-end on one lane of a Illumina GAIIx flow cell at

Barts and the London Genome Centre. The same alignment and SNP calling methods used in phase one was applied to 75 exomes, with SeattleSeq and Annovar variant annotation. Quality control steps applied were a phred quality score of >20 and >8x coverage for each variant position.

3.5.2.1 Sanger Capillary Sequencing

For follow-up confirmation of identified novel variants and segregation analysis, capillary sequencing was applied to all familial cases and controls, where DNA was available. Sanger sequencing was performed on PCR products. PCR primers were designed flanking approximately 200bp of a given variant and outsourced to Source Bioscience for dideoxy sequencing in forward and reverse. Sequencing reads were analyzed using the Sequencher software package (GeneCodes Inc.).

Some exome sequenced individuals had rare novel exome SNP genotypes from the Illumina Immunochip array that allowed allele frequencies to be established in the absence of DNA for Sanger sequencing.

3.5.3 Phase Two: Results

One Illlumina GAIIx 76bp PE lane provided an optimal ~50x mean read-depth (on-target, non-duplicate reads) (Figure 3.3). This read depth was sufficient to call 15-16,000 variants in each sample, and no difference in the number of variant calls was observed at a higher read depth. A high number of SNVs (54,939) were present in 1000G (2011 release) and dbSNP130, but 37.8% (33,323) of these variant calls were novel in 75 exomes, and 5,839 were novel and LoF (Figure 3.4 and Table 3.2). Here, LoF is defined as a mutation that causes reduced or complete loss of protein function such as: frameshift (insertion or deletion of a number of bases that is not a multiple of 3, usually introducing a premature stop codon and lots of amino acid changes), splicing (variant is within 2bp of splice junction), stop loss (SNV that leads to creation of stop codon which can be nonsynonymous or frameshift/nonframeshift indel),

stop gain (SNV that leads to elimination of stop codon) and nonsynonymous nonsense.

Figure 3.3: Number of non-reference SNP calls per exome (n = 75) and corresponding average read depth

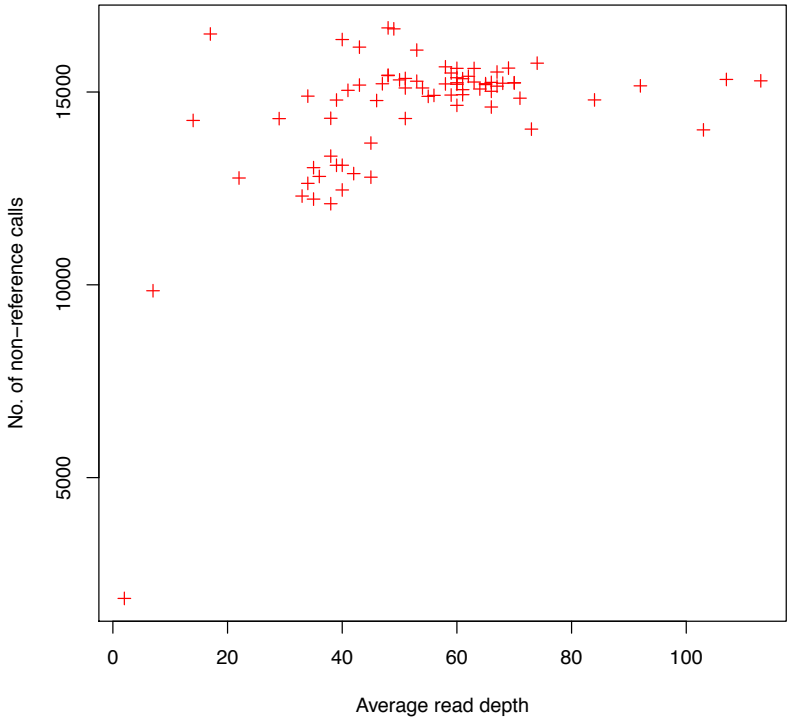


Figure 3.4: Total number of reference and non-reference variants

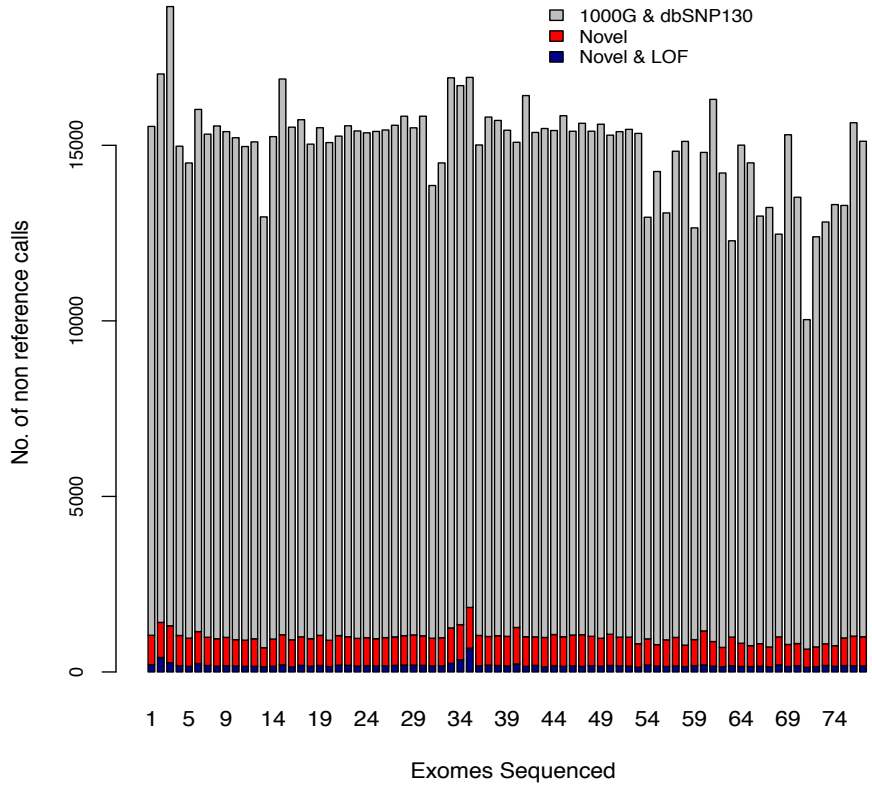


Table 3.2: Total number of SNV calls in 75 exomes

	75 exomes
Total number of unique SNVs	88,262
Novel unique SNVs	33,323
LoF unique SNVs	8114
Novel LoF unique SNVs	5839

Novel defined as not observed in dbSNP130 or 1000G 2011-release reference dataset

Cross validation for SNP calling methods showed high consistency between two samples sequenced at two different sites (Figure 3.5, completed by Ian Stanaway at University of Washington): SAL-12583-9 coeliac sample sequenced here and NA12878 HapMap control sample sequenced at the University of Washington. Comparison of Novoalign read mapper and a custom Bayesian SNP

caller with BWA alignments and Samtools/picard/GATK annotation methods showed near identical outcomes at high read depth, but the former was more sensitive at finding SNPs at a lower coverage. On average, 80% of reads were on target with Novoalign mapped reads. Hap300 GWAS data was available for 26 high quality exomes and 99.96% of exome SNP calls were concordant with Hap300 genotype SNP calls (total of 49,551 heterozygote calls in both datasets and 21 heterozygous calls in sequence but homozygous in GWAS data). Within this set of call positions, the high concordance with array-based genotypes provides an estimate of sensitivity for rare variant detection, as rare variants are largely expected to be heterozygous.

Figure 3.5: Exon library comparison between coeliac SAL-12583-9 and control NA12878 samples with two different calling algorithms

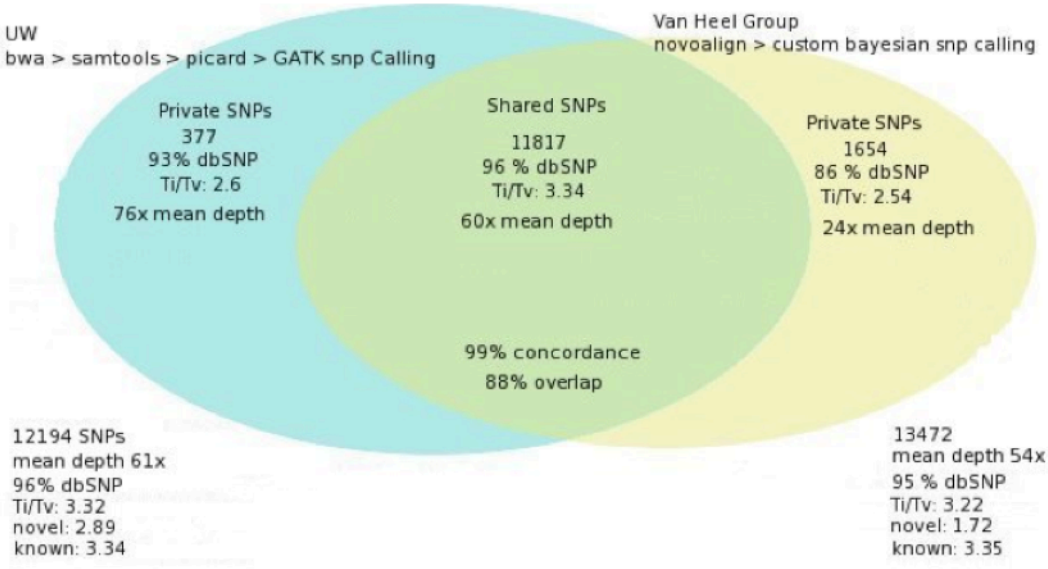


Figure produced by Ian Stanaway at University of Washington.

3.5.3.1 Shared variants between related exomes and segregation analysis

David van Heel performed early 'shared variant' analysis with SeattleSeq annotation software that annotated variants using RefSeq NCBI (build 36), CCDS and 1000G (2010 release) genes and dbSNP130 identifiers. SeattleSeq uses missense and nonsense for nonsynonymous substitution definitions. Variant calls were further annotated based on PolyPhen score, which predicts SNP impact on protein function (Ramensky, Bork et al. 2002) and PhastCons conservation score, a program that detects highly conserved sequences across species and produces a conservation score for candidates that are functionally important (Margulies, Blanchette et al. 2003). To identify rare and low frequency novel SNVs restricted to immune and related disease-pathway genes, in line with expectations from previous coeliac studies, the following filters were applied to heterozygous variant calls and indels: not in dbSNP130, <5% MAF in 1000G 2010 release reference dataset, <10% MAF in coeliac exomes, not in 101 control exomes (54 ultra rare diseases from Kings College London and 47 EGP project samples from Prof. Debbie Nickerson).

Table 3.3 outlines shared mutations in sequenced exomes from the same family and subsequent validation and segregation results. For FAM002, missense mutations in *LGR5*, *SPIC* and *CD180* were confirmed by Sanger sequencing as true positives, however the only informative affected individuals for segregation tests were two children of SAL-12706-6, and this was considered too weakly powered. For SDY, the c.1879G>T nonsense substitution in *IFIHI* was validated in two out of three exomes tested. The entire SDY family had Immunochip genotypes for this SNP so direct segregation was performed on genotypes rather than Sanger sequence data; the substitution was observed in three unaffected and three affected individuals and two affected individuals carried the wild type AA alleles, resulting in no segregation of this variant in familial cases. The missense SNP c.1278G>C in *IL12RB2* and nonsense SNP c.2296C>T in *MADD* shared in FAM006 were validated in all exomes but DNA was not available for any other cases for segregation tests. Additionally, genotype data

was only available for two exome sequenced individuals (SAL-13281-5 and SAL-12575-0) so genotype segregation could not be performed, however, SNP allele frequencies and corresponding *p* values were identified from Immunochip data. For SNPs in *IL12RB2* and *MADD* the association results highlighted no significant p values: *P*=0.9015 (OR 1.006, chi-square (1df) 0.01266, MAF 0.046) and *P*=0.452 (OR 1.053, chi-square 0.5655, MAF 0.064), respectively. For FAM007, high throughput sequencing calls for SNPs in *PTGS2* and *NFIL3* were observed in one direction only, which suggested a false positive call. All SNPs were Sanger sequenced and confirmed false positive. DNA was unavailable for all samples in the DA family so segregation could not be confirmed. From table 3.4, two families were taken forward for segregation analysis.

Table 3.3: Rare nonsynonymous nonsense and missense mutations shared by related coeliac individuals

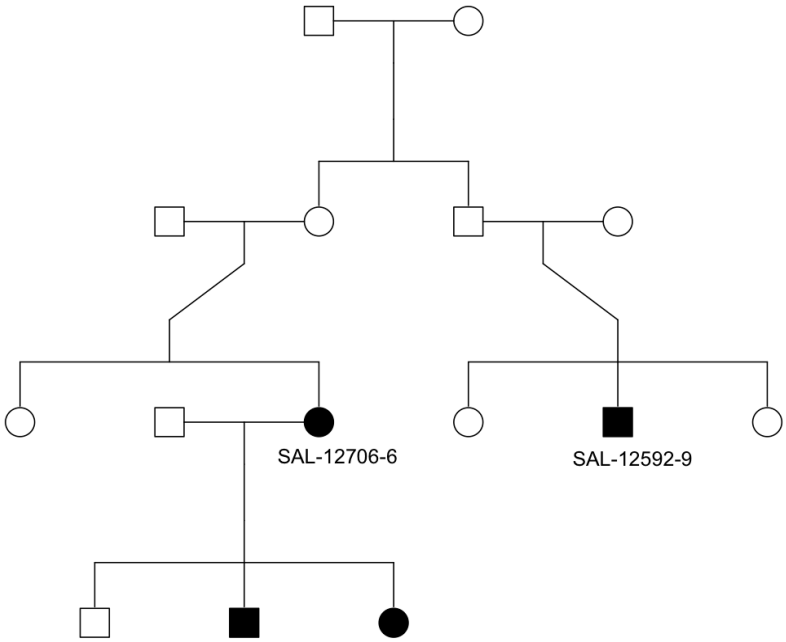
Family ID	Samples sequenced	Relationship	Selected SNPs of interest	Gene	Mutation type	Polyphen prediction	In dbSNP 130 or 1000G ?	PhastCons score	Exomes tested/ Exomes Sanger validated
FAM002	SAL-12592-9 SAL-12706-6	1st cousins	c.1270C>G	*LGR5*	Missense	Benign	No	0.999	2/2
			c.982C>G	*CD180*	Missense	Benign	No	1	2/2
			c.109A>G	*SPIC*	Missense	Benign	Yes	1	2/2
SDY	SDY11 SDY20 SDY101	Uncle to SDY20 and SDY101 1st cousin to SDY101	c.1879G>T	*IFIH1*	Nonsense	Missing domain	dbSNP only	0.988	2/2
FAM006	SAL-13281-5 SAL-12575-0	1st cousins once removed	c.1278G>C	*IL12RB2*	Missense	Possibly damaging	Yes	0.913	2/2
			c.2296C>T	*MADD*	Nonsense	Possibly damaging	Yes	0.941	2/2
FAM007	CAP152916 CAP200010	1st cousins	c.122G>T	*PTGS2*	Nonsense	Probably damaging	No	1	2/0
			c.510G>A	*NFIL3*	Nonsense	Probably damaging	No	0.26	2/0
BRK	BRK4	Grand-uncle to	c.1153G>A	*IL22RA1*	Missense	Benign	No	0	2/1

BRK	BRK11	BRK11	c.1153G>A	*IL22RA1*	Missense	Benign	No	0	2/1
			c.184C>T	***TNFRSF21***	Missense	Benign	No	0.998	2/2
DA	DA269 DA194	1st cousins	c.1421A>T	*PXK*	Missense	Benign	No	0.984	2/2
NEU7017	38481 36790 37456	2nd cousin once removed to 36790 and 37456 1st cousin to 37456	c.358T>A	*CEACAM7*	Nonsense	Probably damaging	No	0	3/0
NEU4801	33210 40123 33165	1st cousin twice removed to 40123 Grand-nephew to 33165 1st cousin to 33210	**c.70G>A**	***IL21R***	Missense	Probably damaging	No	0.155	3/3

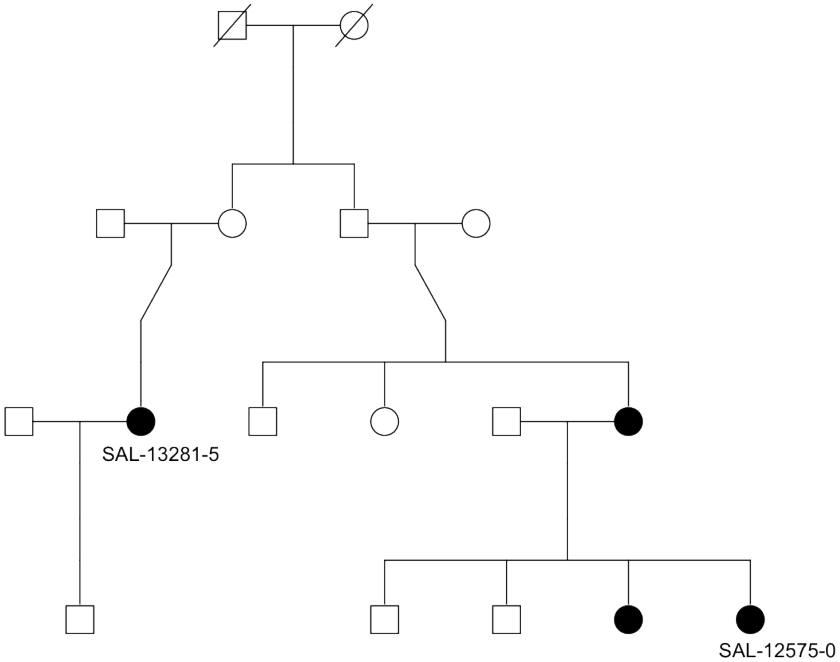
2010 release 1000G dataset and dbSNP130 used for filtering. SNPs in bold were taken forward for segregation analysis.

Figure 3.6: Pedigrees from Table 3.4

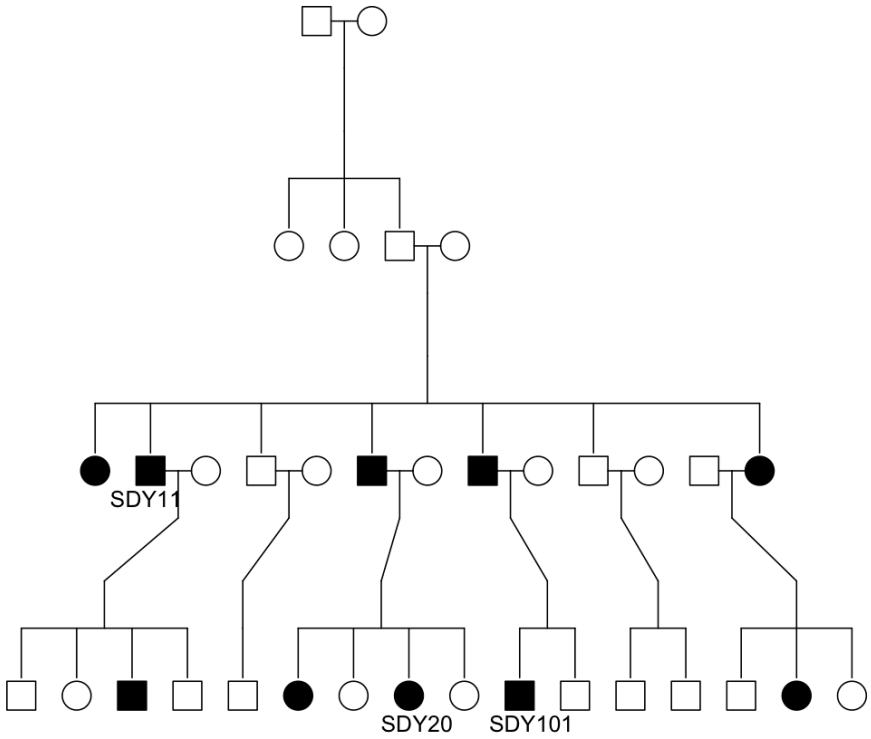
FAM002



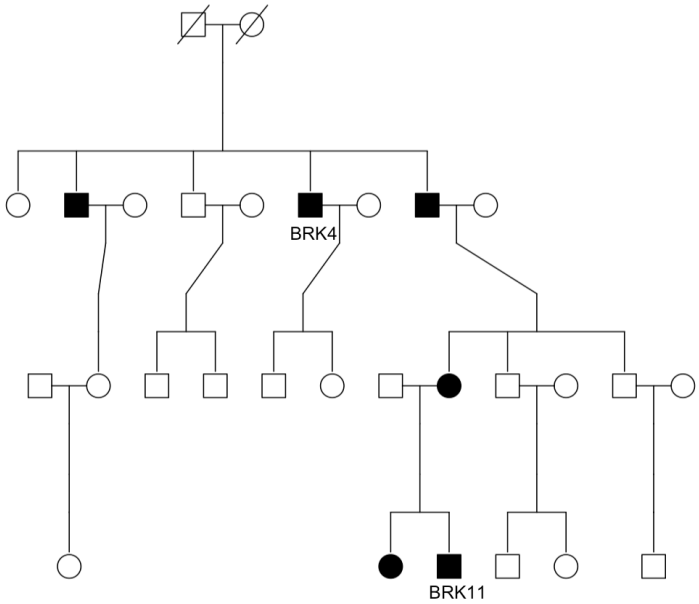
FAM006



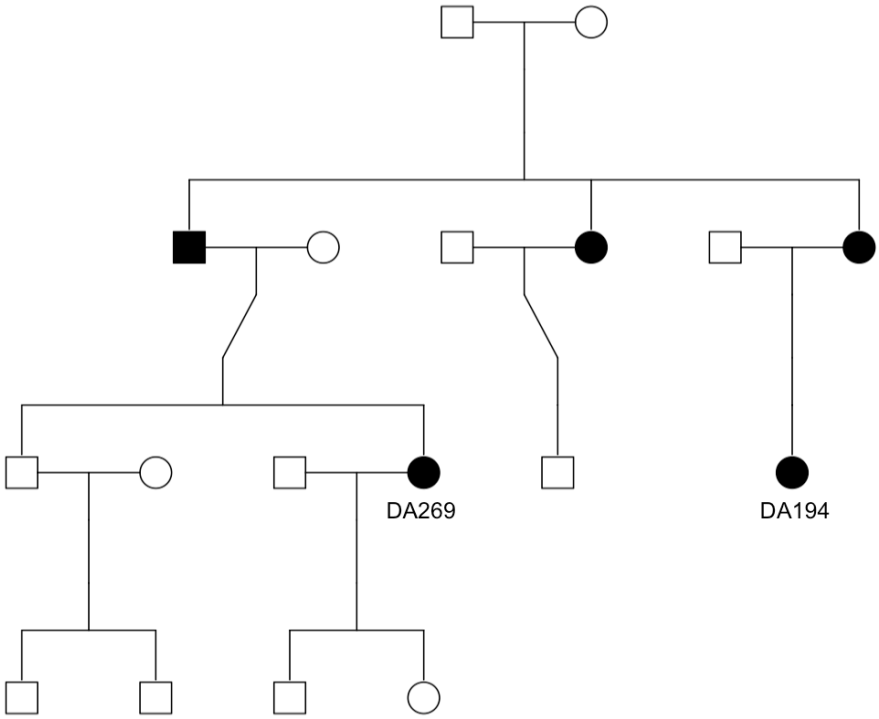
SDY



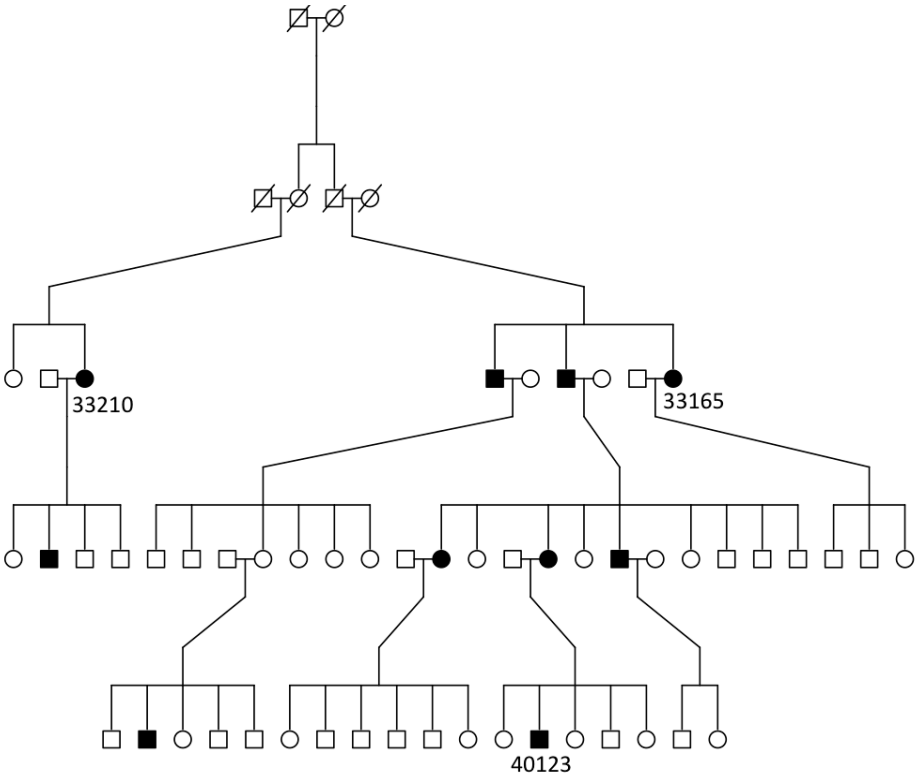
BRK



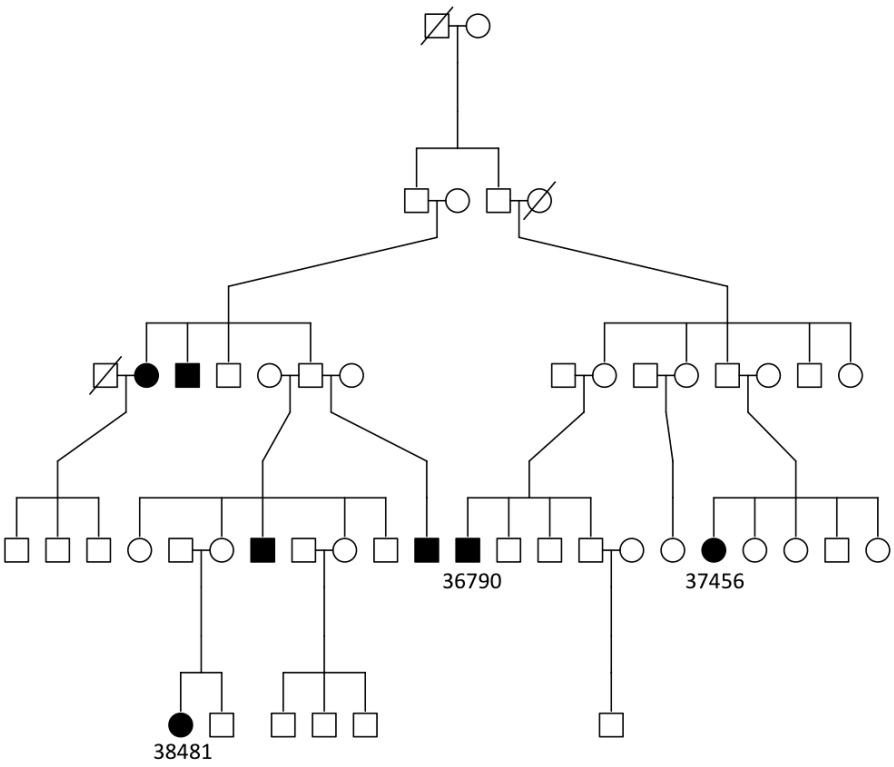
DA



NEU4801



NEU7017

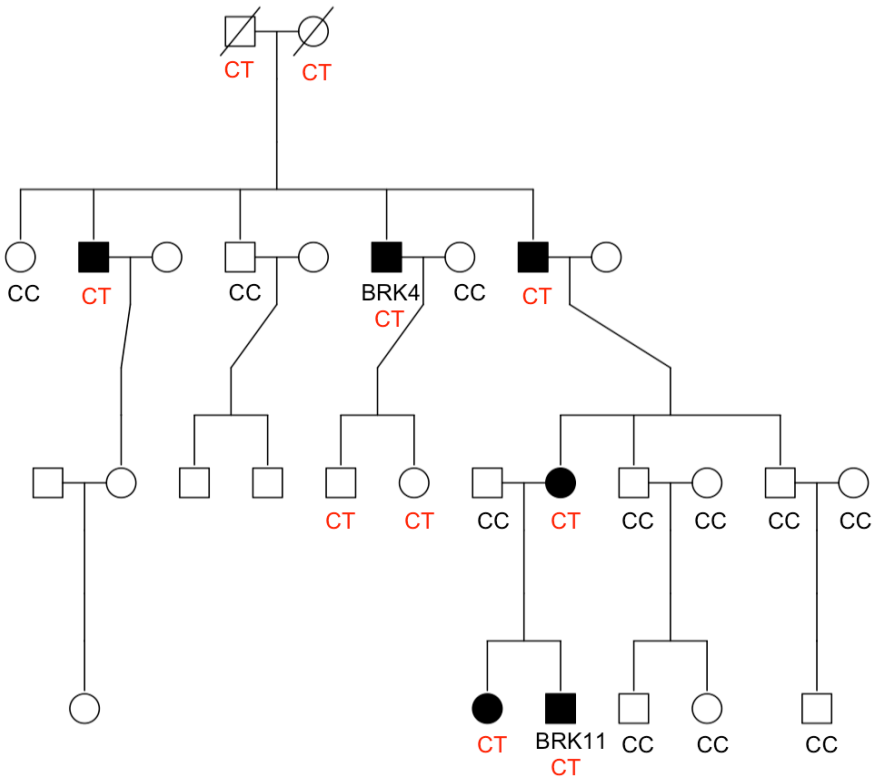


Black = coeliac cases. Pedigrees show relationships of those individuals chosen for exome sequencing. The further apart the relationship, the more likely they are to share a rare disease causing variant in a smaller recombinant region. Sample names for exome sequenced individuals shown only. Family tree for FAM007 was not available because the correspondent did not reply to any email and letters.

TNFRSF21

TNFRSF21, also known as DR6, is down-regulated in active T cells and DR6-deficient mice display reduced *CTLA4* expression, CD4+ T cell proliferation and T-helper cell differentiation (Liu, Na et al. 2001; Zhao, Yan et al. 2001), implicating a possible role in inflammation. Sanger sequencing confirmed the presence of the nonsynonymous missense substitution, c.184C>T (p.G62S), in five of the five coeliac cases and two of the 13 unaffected relatives in the BRK family (Figure 3.7). Two heterozygous genotypes in unaffected individuals suggest a) *TNFRSF21* is necessary but not sufficient for disease development i.e. dominant with reduced penetrance or b) undiagnosed coeliac cases.

Figure 3.7: Segregation result for novel c.184C>T SNV in *TNFRSF21* in BRK family

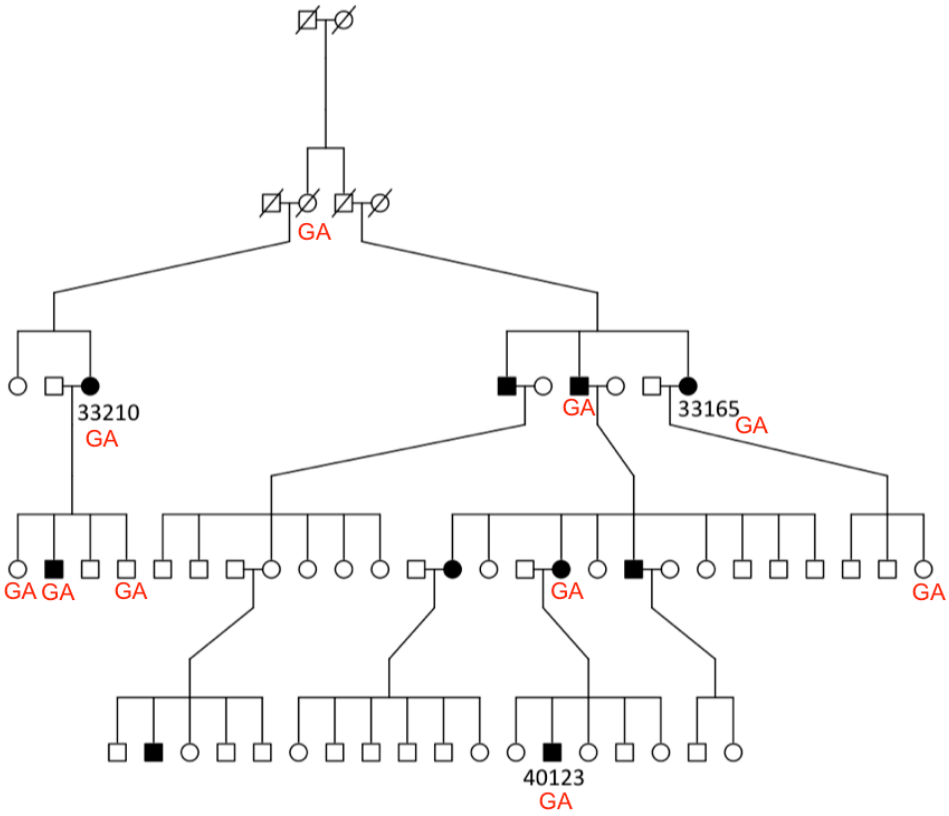


SNP c.184C>T was sequenced in 20 individuals; DNA for 5 members was not available.

IL21R

IL21R, a cytokine receptor for interleukin 21, is selectively expressed in lymphoid tissues and is important for proliferation and differentiation of B cells, T cells and NK cell expansion (Parrish-Novak, Dillon et al. 2000). RA patients also express *IL21R* in macrophages and fibroblasts from synovial joint biopsies (Jungel, Distler et al. 2004). Sanger sequencing identified the nonsynonymous missense substitution, c.70G>A (p.V24I), in six out of the ten coeliac cases and four out of 37 unaffected relatives (Figure 3.9). The wild type alleles, GG, were observed in rest of the family (genotypes not shown in Figure 3.8).

Figure 3.8: Segregation result for novel c.70G>A SNV in *IL21R* in entire Neu4801 family



All other individuals carry homozygous GG wild type alleles

As of June 2010, a new method became available for SNP annotation (Wang, Li et al. 2010), and a second more comprehensive analysis was performed on familial exomes with Annovar annotated SNPs (RefSeq NCBI build 37, 1000G 2011 data release, and dbSNP132 identifiers) to locate more shared variants in immune genes by Vincent Plagnol (Table 3.4). In place of PhastCons, SIFT scores were assigned to heterozygous calls to predict whether an amino acid substitution had any phenotypic effect (Ng and Henikoff 2003). The results were then passed onto myself for validation. Tests for segregation could not be performed in families shown in table 3.5 because DNA on other affected and unaffected members was not available.

In addition to *TNFRSF21*, identified from family segregation, a further two genes in the same gene superfamily were observed in FAM002 and FAM007. *TNFRSF10A* is involved in cell death and apoptosis and a SNP in this gene was associated in cases with age-related macular degeneration with a significant *p* value of 1×10^{-12} (Arakawa, Takahashi et al. 2011). More interestingly, *TNFRSF13* is involved in T-cell independent B cell antibody responses and B-cell homeostasis and has been associated with serum IgG levels in the Chinese population, which has a role in humoral immunity (Liao, Ye et al. 2012). All nonsynonymous substitutions, except those in *CFTR* and *C4PBA* (in one of two individuals), were true positives and possible candidates for resequencing.

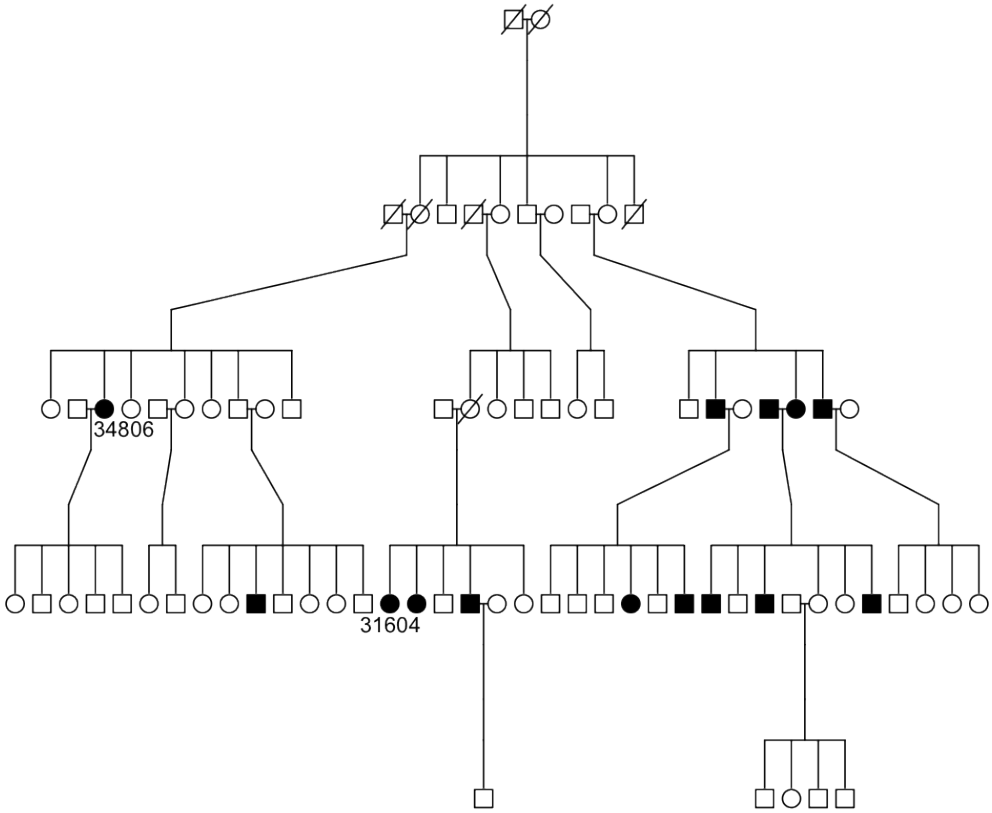
Table 3.4: Rare non-synonymous single nucleotide variants located in immune genes and shared by related coeliac individuals with Annovar annotation

Family ID	Samples sequenced	Relationship	Selected SNPs of interest	Gene	Mutation type	Polyphen prediction	In dbSNP130 or 1000G?	SIFT score	Exomes tested/Exomes Sanger validated
FAM002	SAL-12592-9 SAL-12706-6	1st cousins	c.617G>A	*C4PBA*	Nonsyn	Benign	No	0.24	2/1
			c.58C>T	*TNFRSF13B*	Nonsyn	Probably damaging	No	0	2/2
			c.517G>A	*TRAF4*	Nonsyn	Benign	Yes	0.02	2/2
FAM006	SAL-13281-5 SAL-12575-0	1st cousins once removed	c.66T>G	*RAF1*	Nonsyn	Possibly damaging	No	0.03	2/2
			c.223G>A	*MAP4K2*	Nonsyn	Possibly damaging	No	0	2/2
FAM007	CAP152916 CAP200010	1st cousins	c.1251C>A	*CFTR*	Nonsyn	Benign	dbSNP130	0.57	2/1
			c.1232T>C	*TNFRSF10A*	Nonsyn	Probably damaging	1000G	0	2/2
			c.961C>T	*HAS1*	Nonsyn	Possibly damaging	No	0.01	2/2
NEU4768	31604 34806	1st cousins once removed	c.588A>C	*C1QBP*	Nonsyn	Possibly damaging	No	0.11	2/2
			c.222C>G	*IL12RB1*	Nonsyn	Possibly damaging	dbSNP130	0	2/2

SIFT scores range from 0 to 1, where <= 0.05 is predicted damaging and >0.05 is predicted tolerant

Figure 3.9: Pedigrees from Table 3.5

NEU4768



Pedigrees for FAM002 and FAM006 are illustrated in Figure 3.6. Pedigree for FAM007 was not available.

3.5.3.2 Single-SNP and aggregate tests for rare variants

Analysis in the study so far has based criteria for variant searching by the filtering approach and largely focused on loss of function variants only. A second more practical approach taking into account all sequenced variants was used to perform single SNP and gene-level burden tests. The single SNP test compared calls from 222 neurological disorder control exomes (captured with Agilent Sure Select version 1) to coeliac exomes, similar to the test one would apply in a GWAS. An excess of rare variants in the HLA-complex on chr6 was observed, with significant *p* values ranging from 10^{-4} and 10^{-7}, as illustrated in the

Manhattan plot (Figure 3.10). No other SNP reached $p=10^{-7}$ or higher. A synonymous SNP in *NDUFV2* on chr18 reached $p=1.26^{-6}$ (MAF 0.0187); mutations in this gene are associated with Parkinson's disease (Hattori, Yoshino et al. 1998) and bipolar disorder (Washizuka, Kakiuchi et al. 2003; Doyle, Dahl et al. 2011) highlighting that the signal is likely to be associated with one of the neurological diseases in the control exomes rather than CD. While the test accounted for target capture efficiency and only calls with comparable call rates were used, there are still evident pitfalls using different capture platforms (Agilent Sure Select for controls compared to Roche NimbleGen for coeliacs) and likely false positives were evident (see quantile-quantilte (Q-Q) plot in Figure 3.11).

Figure 3.10: Manhattan plot of single-SNP tests comparing the case data (n = 41) with the control samples (n = 222)

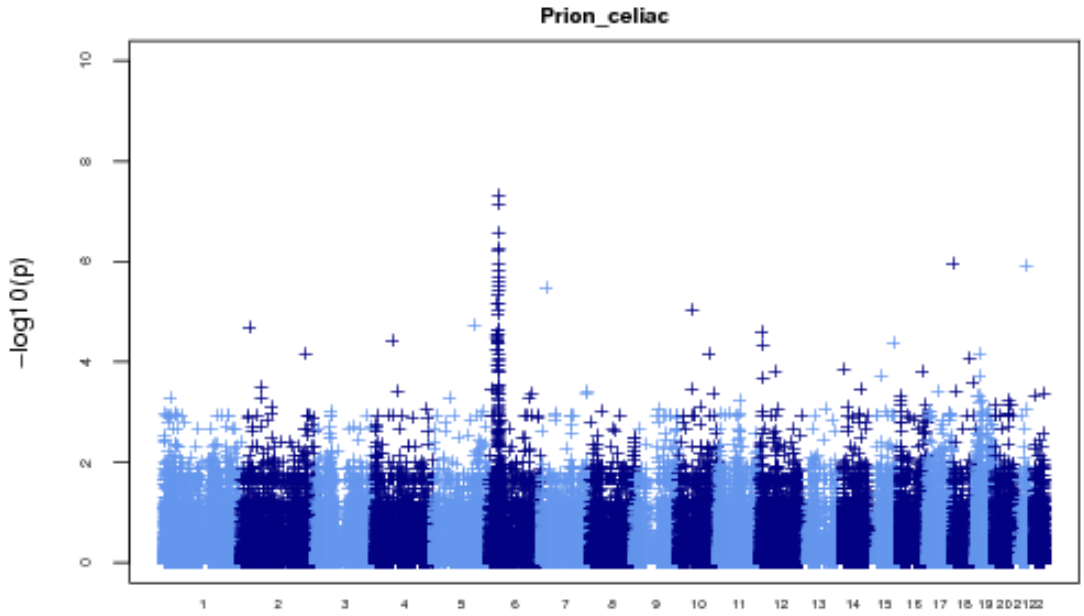
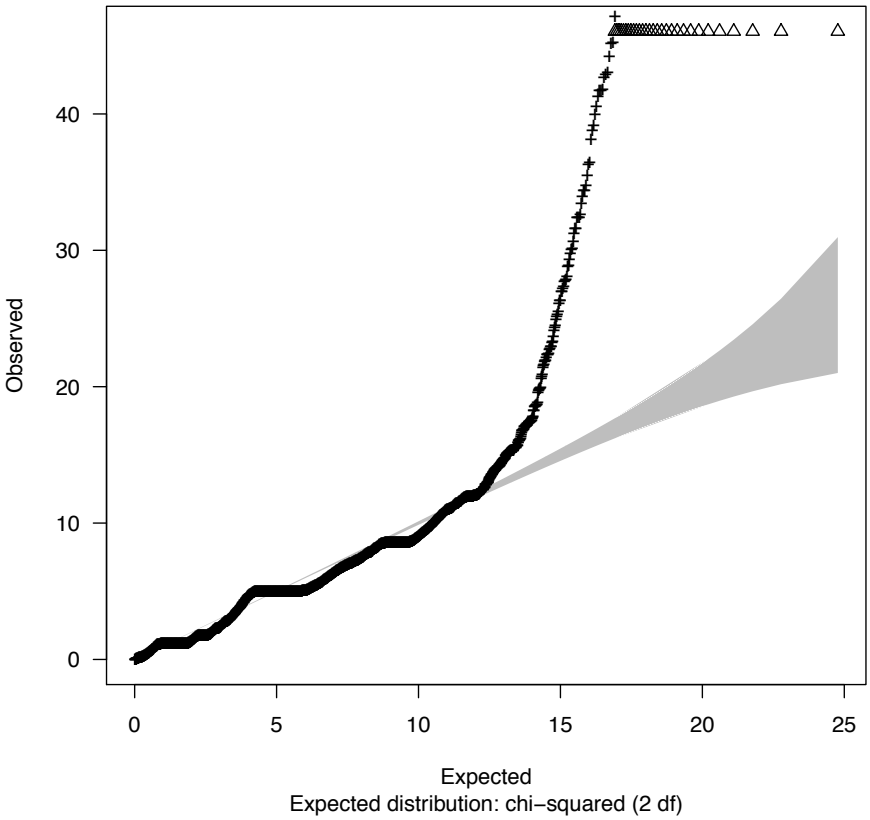


Figure 3.11: Q-Q plot of single-SNP tests comparing case data (n=41) with control samples (n=222)



An aggregate test for rare variants in a complex trait, using a minor allele frequency based on 1000G, offers a genome wide approach that limits problems that can be associated with SNP filtering: within-gene heterogeneity and reduced penetrance. This type of test compares the number of variants within a gene to the genome-wide distribution of rare variants in the same functional category to derive a gene-based Fisher exact *P*-value (two-tailed) (Stitziel, Kiezun et al. 2011; Kiezun, Garimella et al. 2012). The test aggregates variants into discrete features (a natural grouping unit in the exome is a gene) to obtain greater statistical power. This is achieved by reducing multiple tests, as the number of genes containing aggregated rare variants is tested rather than one test per rare variant, and combining allele frequencies of aggregated variants to achieve a higher overall allele frequency compared to small individual rare variant allele frequencies.

Three tests were performed comparing SNP calls in cases and controls based on a genome-wide distribution of rare alleles. A single-SNP *P*-value from multiple variants in a gene was combined and derived from a two-tailed Fisher exact test, allowing the same inferences as one would make in a genome wide association test. Related exome-sequenced individuals were removed to eliminate bias, and the remaining 41 exomes were compared to 222 neurological control exomes; only variants with a MAF <0.5% in 1000G 2011 reference dataset were observed. Genes with rare variants in all deleterious functional categories are shown in Table 3.5. Table 3.6 lists genes harbouring loss of function variants only. Table 3.7 lists genes with loss of function variants in immune genes.

Table 3.5: Top 5 most significant genes for the aggregate test rare variants (LoF, non-synonymous and splice site) between cases and controls

Gene	Number of rare alleles in controls (n = 222)	Number of rare alleles in cases (n = 41*)	Fisher *p*
PER2	6	9	0.00057
PLEKHA6	0	4	0.00097
FLG	5	7	0.0026
SLC3A1	2	5	0.0029
WDR59	2	5	0.0029

*Total number of cases after removing additional exomes within each family. Rare allele is defined by a frequency less than 0.5% in the 1,000 genomes data (n = 1092).

Table 3.6: Top 3 most significant genes for the aggregate test for rare LoF variants only between cases and controls

Gene	Number of rare alleles in controls (n = 222)	Number of rare alleles in cases (n = 41*)	Fisher *p*
ITGAE	0	2	0.027
TEX14	0	2	0.027
CUBN	2	3	0.043

*Total number of cases after removing additional exomes within each family. Rare allele is defined by a frequency less than 0.5% in the 1,000 genomes data (n = 1092).

Table 3.7: Top 15 most significant genes for the aggregate test for rare LoF variants in immune genes between cases and controls

Gene	Number of rare alleles in controls (n = 222)	Number of rare alleles in cases (n = 41*)	Fisher *p*
CD1C	0	3	0.005
CERK	0	3	0.005
CRLF3	0	3	0.005
DDR1	2	4	0.010
HLA-DOA	4	5	0.012
ZFYVE16	4	5	0.012
IKZF3	1	3	0.016
RPS6KA2	1	3	0.016
CDH17	3	4	0.020
LPP	5	5	0.020
CD180	0	2	0.022
CTGF	0	2	0.022
DNM1L	0	2	0.022
EB13	0	2	0.022
IFNW1	0	2	0.022

*Total number of cases after removing additional exomes within each family. Rare allele is defined by a frequency less than 0.5% in the 1,000 genomes data (n = 1092).
The results in tables 3.6 and 3.7 are based on multiple testing corrections hence the observed differences in *P* values; the test in table 3.7 contained a lower number of genes than the test in table 3.6, so the penalty for multiple testing was reduced.

Genes in table 3.5 did not appear to have any potential function for CD susceptibility, or any other overlapping disease where one can deduce a shared function. For example, an excess of rare variants in cases and controls in *PER2* is possibly owing to its function as a circadian pacemaker in the mammalian brain involved in behavioral and metabolic factors, rather than being enriched for CD risk variants; mutations in *SLC3A1* are associated with cystinuria, an autosomal recessive disease characterized by kidney stones (Pras, Raben et al. 1995); *FLG*, a gene that encodes the filaggrin protein that forms a component of the skin barrier, has strongly associated LoF variants in atopic eczema and ichthyosis vulgaris (Sandilands, Terron-Kwiatkowski et al. 2007) but no association has been implicated in InBD susceptibility (Van Limbergen, Russell et al. 2009).

Based on protein function, *ITGAE* and *CUBN* were suggestive candidates for further screening. *ITGAE*, also known as *CD103*, encodes an alpha integrin involved in tissue specific retention of T lymphocytes at the basolateral surface of intestinal epithelial cells and is a possible accessory function for the activation of epithelial cells (Cepek, Parker et al. 1993; Sheridan and Lefrancois 2011). Two confirmed novel stop gain (nonsense) SNVs in *ITGAE* c.2962G>T (p.Glu988X) (identified in SAL-12553-6 from FAM014) and c.314T>A (p.Leu105X) (identified in Neu7058-39198 from Neu7058), were not present in 222 controls. Both SNPs were tested for segregation in all affected and unaffected individuals of FAM014 and Neu7058. The c.314T>A substitution was present in four individuals in Neu7058, three of which were non-disease cases. Only one other unaffected individual carried the c.2962G>T substitution in FAM014. Neither mutation segregated with disease in two families.

CUBN (cubilin) is located on chromosome 10p21.1 and is expressed within the epithelium of the intestine where it acts as a receptor for intrinsic factor-vitamin B (12) complexes (Fodinger, Wagner et al. 2001). Missense and insertion mutations in this gene have been associated with megalobastic anaemia in Finnish families (Aminoff, Carter et al. 1999), a rare autosomal recessive condition characterized by selective intestinal vitamin B12 malabsorption. It is not known whether the three individuals bearing the nonsense mutation in this gene have megaloblastic anaemia; it is common for

CD patients to have low B12 and folate levels, causing pernicious anaemia. A recent meta-analysis to identify risk variants for albuminuria for early prevention of chronic kidney disease located a risk variant in *CUBN* to be associated with albuminuria level in individuals with diabetes (Boger, Chen et al. 2011). Three novel stop gain (or nonsense) mutations in *CUBN* (RefSeq accession number NM_001081) were observed in three separate individuals: c.4459C>T (p.Arg1487X), c.5428C>T (p.Arg1810X) and c.6359G>A (p.Trp2120X). All substitutions are possibly damaging, predicted by PolyPhen and GAII*x* sequence pile-up data indicated real heterozygotes with a high read depth (173, 44 and 53 respectively), confirmed by Sanger sequencing.

Overall, candidate genes harbouring true (as confirmed by Sanger sequencing) rare variants, i) shared by related exomes, ii) that showed a higher burden in cases than controls, and iii) that segregated in familial disease cases, were selected for resequencing based on interesting immune function, size and number of exons.

3.6 Chapter discussion

Strategies to discover rare major impact variants in common disease have been widely discussed (Cirulli and Goldstein 2010; Eichler, Flint et al. 2010) and exome-sequencing based studies are a popular approach to test for association of rare coding variants with complex phenotypes. The empirical successes of candidate gene resequencing (Ji, Foo et al. 2008; Johansen, Wang et al. 2010) and Mendelian studies suggest a large portion of disease-associated variation lie within coding exons (Cooper, Ball et al. 1998; Botstein and Risch 2003; Glazov, Zankl et al. 2011). Based on this, it was likely that many rare mutations in a gene(s) were to be located that could contribute to missing disease heritability.

The 75 coeliac sample dataset contained an abundance of rare coding variants (~33,000) and sequencing additional samples would probably continue to reveal additional rare variants. Keizun et al. discovered that as sample size increases the number of observed variants increases (an average of 40 times more

nonsense variants in 300 samples than in a single sample alone) owing to purifying selection and recent population expansion (Kiezun, Garimella et al. 2012). Large characterized families are therefore as important in exome sequencing as in positional cloning or linkage methods, not only for variant enrichment but to filter potential risk variants. Protection from confounding factors owing to population stratification and allelic heterogeneity is another advantage of family designs. DNA-enrichment methods and massively parallel high throughput sequencing promise high quality data of captured protein-coding variants - of these functionally important variants, any two shared between related individuals are likely to be causal if the variant segregates with disease. If a rare variant of large effect was to be functionally defective in a family, LoF, splice site and nonsyonymous mutations are a clear starting point for analysis. This approach was adopted to search for a Mendelian or near Mendelian form of CD. The second approach was to implement a case-control method by performing single-SNP tests and rare variant gene-burden tests to search for mutations in genes reaching significance. The focus on immune genes extended from GWAS and fine mapping results in CD. The latest fine mapping of current coeliac loci identified 13 new coeliac risk loci at genome-wide significance (Trynka et al. 2011), bringing the total to 40 (including the HLA). Over 90% of associated loci are represented in known immune system genes so it was hypothesised that any rare variants of large effect would compromise immunological function.

Specificity of variant identification was assessed based on concordance rates with genotyping data and the percentage of Sanger sequencing false positives. Here, a false positive is defined as the appearance of an allele at a sequenced locus in the NGS data that is not observed in the Sanger sequencing data. High specificity of variants in the sequencing dataset was observed by >99% correlation with a GWAS common SNP HapMap control. Overall, 37 SNPs were Sanger sequenced across all exome individuals. Three of the 37 were complete false positives in the NGS data, resulting in an 8.1% false positive rate. Four out of 37 SNPs were false positive in only one of the two related exomes, possibly due to low coverage at that variant site for the second individual leading to a

sequencing error. Sequencing error decreased with better variant annotation software; there were three complete false positives with SeattleSeq annotation compared to zero with Annovar annotation. Furthermore, the top 400 novel exome SNPs on the Immunochip array (discussed further in Chapter 4) have a false positive rate of 25%; this error estimate comprises not only sequencing miscalls, but clustering miscalls and any assay fails. Despite stringent quality control error can still be present in sequencing data and unfortunately this is the case when stratifying variants on the basis of putative functional consequences; the class of variation that is annotated to be most deleterious is also more heavily enriched for errors (MacArthur and Tyler-Smith 2010). Having an up-to-date reference dataset, such as the 1000 Genomes, helps identify true LoF variants.

The first large-scale exome resequencing study was by Li et al. who found an excess of low frequency rare mutations in 200 exomes (Li, Vinckenbosch et al. 2010). This data was not used as a permanent filter in the coeliac dataset because of early technology and limited coverage. Overall, all 75 sequenced coeliac exomes passed quality control with >13,000 SNP calls at an average read depth of 50x per exome, compared to 12x in the Li et al. dataset. One lane of 76bp paired end data (v4 chemistry, GAII*x*) typically provided ~21 million on target non-duplicate reads. The in-solution DNA capture method proved markedly superior to array based capture and was optimized to decrease duplicate reads (reduction of PCR cycles) and increase data coverage for variant detection (no multiplexing). Yet, across all 75 exomes, an average 80.4% of post-capture non-duplicate on target reads was obtained (Appendix I-C), meaning 20%, or 4-8 Mb (1000 – 2000 genes), of the target region was not covered with sufficient read depth for variant detection. As far as sequencing goes, these un-captured regions will not be observed in the dataset. However, even within the set of called variants, there is also an issue of whether a true heterozygote in captured regions is in fact a heterozygote variant. The same number of reads per allele is required to call an heterozygote but if alleles are captured at different rates in the library processing step, then a sequencing miscall is possible. Sanger sequencing validation is therefore a necessity with

sequencing methods to validate true calls. Regional variability is also a concern across different target capture methods for exome sequencing. Some regions can have greater technical artefacts due to overlapping probes, or may be under-covered to high GC content or segmental duplications preventing accurate alignment to the reference sequence (Hedges, Guettouche et al. 2011). There is also a problem of variable uniformity in NGS platforms. A mean coverage does not mean that each base was read the reported number of times; some are read only occasionally whilst others are oversampled. Further sequencing will increase the number of unrepresented reads, but array-based genotyping to utilize maximum variant calling is also an option.

The key challenge for disease gene discovery using exome sequencing is the huge number of apparently private mutations present by chance in any single genome, making it difficult to decide which variant is causal, even if only nonsynonymous, splice site and indel mutations are considered. This early observation was highlighted in the phase two dataset, where thousands of protein coding mutations per individual were identified e.g. 16,044 exonic SNPs were observed in sample CAP200010, of which 6,600 were missense or nonsense and 15,855 exonic SNPs were observed in CAP152916, of which 6,759 were missense or nonsense. This posed a 'needle in a haystack' problem. Having the Immunochip case control dataset helped familial analysis to validate candidate mutations. Two rare SNPs in *MADD* (MAP-kinase activating death domain, T cell expressed) and *IL12RB2* (interleukin 12 receptor) were on the Immunochip assay, but did not reach significance at $P<10^{-5}$. This result expressed the need for high quality control exome data, which came at a later date (222 neurological disease control exomes). Differences in in-solution capture methods used in this study have been evaluated in a prior study (Agilent Sure Select for 222 control samples and Roche NimbleGen for 75 CD exome samples); more high quality reads aligned to target regions with NimbleGen's capture probes (Sulonen, Ellonen et al. 2011). These differences may have had slight impact on the results for the single SNP and gene burden tests.

Although the familial design was adopted in order to filter potential risk variants, segregation tests were either inconclusive due to small sample size or failed to segregate directly in CD cases; the reference wild type alleles were always present in affected cases. One positive segregation test was in the BRK family; all CD cases in BRK carried the nonsense c.184C>T substitution in *TNFRSF21*. Additional CD GWAS-associations genes from the same super-family (*TNFRSF18* and *TNFRSF14*) highlighted *TNFRSF21* as a candidate for resequencing. Surprisingly, for the single-SNP test, the HLA region on chromosome 6 showed strong significance, even with a small sample size (44 cases), but no other significant variant was established. Here the sample size resulted in lack of power, and limited the extrapolation of any true result; the most significant hits were probably sequencing artefact. For example, one would need 200 cases (and 200 controls) to give 0.2% power for a 0.2% frequency SNP to be detected at 10^{-8} (Purcell, Cherny et al. 2003).

The overall aim of analysis from sections 3.5.3.1 and 3.5.3.2 was to identify candidate genes for targeted resequencing to assess if there were a significant burden of rare variants in captured genes from enriched disease samples (Chapter 5). Cost and amplicon design considerations had to be taken into account for the follow-up study, so genes were selected based on likely function in CD susceptibility, size of coding region and number of overall PCR amplicon targets. Despite there being rare novel nonsense variants *CUBN* in coeliac cases, these variants also occur in controls. With 67 exons and 3,623 amino acids, *CUBN* has a relatively large coding sequencing, so majority of the population will have at least one rare (defined as a mean allele frequency <0.5%) missense or truncating variant at this locus. Even in all 75 exome cases, there are an excess of LoF variants in *CUBN* (494), compared to 11 in *CRLF3*, a much smaller gene with 8 exons and 442 amino acids. NGS of the entire coding regions of these genes has the power to define which variants are pathogenic, however when relying on an economy of scale, for a similar cost as sequencing *CUBN* alone, five genes of ~2,200bp in coding length can be sequenced, therefore CUBN was not selected for resequencing. All other candidate genes based on case control rare allele tests and whether shared with another exome are listed in Table 3.8.

Table 3.8: Candidate genes from exome sequencing analyses selected for targeted gene resequencing

Gene	Analysis type	cDNA size (bp)	No. of exons	Known immune?	Validated true positive?
CD1C	Case control	1,435	6	Yes	Yes
CERK	Case control	4,450	13	Yes	Yes
CRLF3	Case control	2,917	8	Yes	Yes
IKZF3	Case control	9,667	8	Yes	Yes
CD180	Case control/shared	2,726	3	Yes	Yes
EBI3	Case control	1,128	5	Yes	Yes
IFNW1	Case control	1,514	1	Yes	Yes
RAF1	Shared	3,300	17	Yes	Yes
TNFRSF10A	Shared	1,357	5	Yes	Yes
MAP4K2	Shared	2,955	32	Yes	Yes
C1QBP	Shared	1,169	6	Yes	Yes
TNFRSF13B	Shared	1,357	5	Yes	Yes
TRAF4	Shared	2,921	7	Yes	Yes
IL12RB1	Shared	2,100	17	Yes	Yes
HAS1	Shared	2,087	5	Yes	Yes
TNFRSF21	Segregation	3,595	6	Yes	Yes

3.7 Chapter Conclusions

The points below conclude the findings from the research in this chapter:

1. Many rare and low frequency (between 0.09% and 5% allele frequencies) SNVs were found in 75 coeliac exomes.
2. No conclusive variants were found to be segregating with disease in large (>2 generation) multiply affected coeliac families, where two or three exomes were sequenced per family.
3. Different analytical strategies were applied to locate genes with potential CD association, within families and across the entire 75-disease case dataset.
4. *CUBN* stood out as a clear candidate given its role as a vitamin B12 receptor and positive Sanger sequence validation, but was too large (11,949bp) for resequencing. Another PhD student may study the role of this gene in CD.

5. 16 candidate genes were sequenced in 2,304 cases and 2,304 controls in a follow-up study. This experiment is outlined in Chapter 5 of the thesis.

Chapter 4

Illumina Immunochip: Linkage analysis, exome SNP case-control association and current coeliac loci contribution in disease cases

4.1 Introduction

The Immunochip array was designed by a cohort of researchers, initiated by the WTCCC, involved in autoimmune genetic studies with a main focus of dissecting rare and common genetic variation at immune regions in autoimmune disease. The custom Illumina Infinium HD array became commercially available in 2010 to Immunochip consortium researchers and contained 196,524 polymorphisms (718 small indels and 195,806 SNPs) across 186 disease susceptibility loci in nine autoimmune diseases (Greco, Corazza et al. 1998; Murray, Moore et al. 2007; Cortes and Brown 2011). Various laboratories contributed data from GWAS significant (P<5 x 10^{-8}) regions and any fine mapping/resequencing regions - the entire SNP content was not genome wide but SNPs from several auto and chronic immune mediated disease loci. This allowed deep replication of GWAS data covering strong candidate genes to dissect true risk signals from top-ranked associations and greater refinement of disease-associated variants that was not possible on the Illumina Hap550 or Omni2.5 chips. Along with established disease loci, all known dbSNP and 1000G (2010 release) variants within 0.1cM (HapMap3 CEU) recombination blocks around each GWAS lead marker on the chip enabled fine mapping of loci for rare and common variants (Polychronakos 2011). A collection of high density mapping studies have been successfully published, using the standard single disease case control approach (Cooper, Simmonds et al. 2012; Eyre, Bowes et al. 2012; Liu, Almarri et al. 2012) and meta analysis with previous GWAS data and imputation (Eyre, Bowes et al. 2012; Jostins, Ripke et al. 2012; Juran, Hirschfield et al. 2012; Tsoi, Spain et al. 2012).

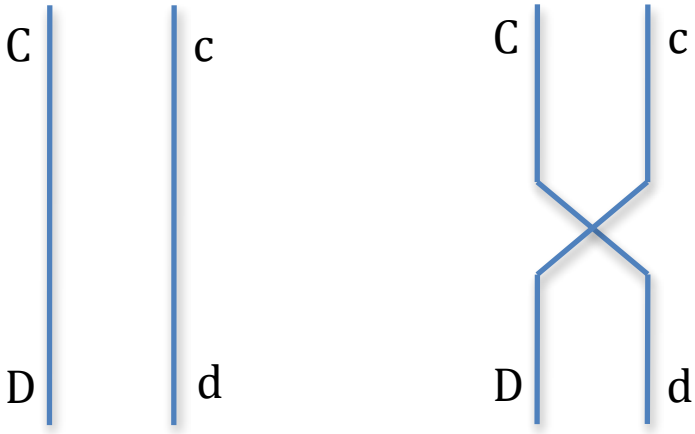
All CD GWAS-associated loci plus a subset of target captured exonic SNPs selected from the exome sequencing dataset in Chapter 3 were contributed by David van Heel (member of the Immunochip consortium) for the Immunochip assay. The CD Immunochip fine mapping study published in 2011 implicated 56 non-HLA SNPs in 40 independent loci to be associated with CD disease risk, meeting the strict criteria of a nominal P value <5 x 10^{-8}, set for large-scale GWAS and fine mapping studies (Trynka, Hunt et al. 2011). This entire chapter

details three sets of analysis using SNP markers from the Immunochip array: i) linkage analysis in coeliac pedigrees using the entire Immunochip marker set (discussed in next sections), ii) case control association analysis of the subset of target captured exonic SNPs selected from the exome sequencing dataset in Chapter 3, iii) genetic load scoring of current 57 (including the highest associated SNP at HLA-DQ2.5) coeliac associated SNPs in coeliac individuals from multiply affected families compared to population controls.

4.1.1 Principles of genetic linkage

The process of recombination is the objective behind linkage inference in family studies. During meiosis, crossing over occurring between two non-sister chromatids is given the term recombination. The resultant homologous chromosomes will have new combinations, which in turn give rise to genetic variation. The amount of crossing over that can take place is dependent on the distance between two genes along the chromosome. Crossover and non-crossover events occur at equal frequency if two genes are far apart; this is the opposite of genes that are closer together, which undergo less recombination. If two loci are linked, the probability that that they are passed down together as a haplotype depends upon the probability of recombination during meiosis, and this probability is translated into a recombination frequency, θ, to determine allele segregation in a disease trait. For example, in figure 5.1, if D is an unknown disease gene and C is an observed marker locus, and if C and D are linked, the marker will segregate in the same way as disease.

Figure 4.1: Linkage and disease genes



Haplotypes: **CD** or **cd** Haplotypes: **Cd** or **cD**

A θ of 50% specifies unlinked loci distantly located on the same chromosome or on separate chromosomes, a consequence of independent assortment. A θ less than 50% means loci are linked with a small physical distance between them.

4.1.2 Linkage models in complex disease

With genetically complex traits, there is unclear evidence of a single locus effect with a specific mode of inheritance, such that the unreliability of conventional linkage analysis requires distinct, model free strategies to be created (Curtis and Sham 1995). For a dichotomous trait genetic complexity is causative of heterogeneity, involvement of multiple genetic loci and environmental exposures, so requires a method where the precise mechanism of disease can remain unknown. Affected sibling pair methods were modelled based on this, followed by likelihood-based methods in which the observed data is a probability of θ of two loci by way of reporting a logarithm (base 10) of odds (LOD) score. The LOD score method was first defined in parametric models where allele frequencies, penetrances and inter-marker genetic distances (for multipoint analyses) were required (Kruglyak, Daly et al. 1996). This model is successful where a clear inheritance pattern is observed, for example, autosomal dominant or recessive. A non-parametric model (NPL, or model free)

is used for disorders of unclear inheritance and variable penetrance. It assumes that linked loci shared by affected relative pairs share more alleles identical by descent (IBD) at the marker locus than expected by chance only (Risch 1990). The test looks for chromosomal segments shared between affected individuals by distinguishing IBD alleles from identical by state (IBS) alleles. Alleles IBS may look identical but their common ancestor cannot be demonstrated, whereas IBD alleles can be (Ott and Bhat 1999). Only genotypic information from the affected individuals is used for analysis, following the assumption that the affected phenotype is more likely to be associated with the presence of the disease allele. Unaffected individuals are used to provide genotypic information on any un-typed parents. Pedigree relationships, allele frequencies and genotypes of all individuals are needed for NPL.

The linkage experiment in this chapter differs to other combined exome/whole genome sequencing and linkage studies in monogenic disease. For monogenic/Mendelian diseases, researchers have combined exome sequencing in pedigrees where positive linkage has been identified to find the causal gene (Louis-Dit-Picard, Barc et al. 2012), and this has been successful for many autosomal dominant (Johnson, Mandrioli et al. 2010; Wang, Yang et al. 2010), recessive (Bilguvar, Ozturk et al. 2010) and quantitative traits (Bowden, An et al. 2010). Linkage has also been used in complex traits. Recently, rare missense mutations in *CARD14* have been implicated in psoriasis, using a variety of strategies, some similar to ones taken in this study. Linkage was identified in PSORS2 chromosomal region, containing *CARD14*. Target capture and sequencing then identified gain-of-function mutations in *CARD14* segregating with disease (Bertin, Wang et al. 2001; Jordan, Cao et al. 2012). For a complete mutation profile of the gene, exon 4 of *CARD14* (where most clustering rare variants were located) was sequenced in 1,856 cases and 882 controls of European ancestry and the entire case (6,000) control (4,000) cohort was genotyped to establish allele frequencies of all rare *CARD14* variants. This study was successful in finding a potential causal gene through linkage, establishing that mutations in the gene also occurred outside of families with a genetic predisposition and that harbouring rare variants are likely to confer a high risk

for the phenotype (Jordan, Cao et al. 2012). The study is a successful one for complex disease and mirrors a monogenic disease finding in that one gene (others genes also contribute to psoriasis susceptibility) containing rare variants are possibly detrimental to disease risk. Furthermore, monogenic forms of complex disease, such as in cutaneous lupus erythematosus, where an autosomal dominant inheritance pattern was described by findings from a genome-wide linkage search in a large kindred (Lee-Kirsch, Gong et al. 2006) provides further evidence of how linkage can identify a single gene responsible for disease susceptibility.

In this chapter, linkage has been performed in large multigenerational coeliac pedigrees using a set of common and rare SNP markers from the Immunochip array. In order the link this information with the exome dataset of 75 CD subjects, genomic regions under the linkage peaks were then inspected for rare exome variants from the exome sequencing dataset (data from Chapter 3, phase two), in the hope of finding segregating variants in one or more genes to take forward for targeted resequencing.

4.2 Aims and hypotheses

Specific aims and hypothesis for three analyses are outlined below:

i) **Linkage analysis** – the aim is to perform NPL analysis on multiply affected families using 196,524 Immunochip SNP markers, to infer if a rare coding disease risk variant is shared by affected individuals in the same pedigree under a linkage peak. The linkage information here provides knowledge of shared chromosomal regions, which can be linked to exome sequencing data to search for rare segregating variants carrying a disease risk. The hypothesis is if common SNPs are segregating more than expected by chance then there will be some rare functional variants under that peak also; the rare variant will be IBD in affected individuals.

ii) **Case control association analysis** - the aim of having exome SNP content on the Immunochip is to observe whether any rare (MAF <0.5% as defined by Immunochip) or low frequency (MAF 1% - 0.5%) mutations captured from 60 coeliac exomes are associated with CD in a large case control dataset. Not only does a large sample size increase power for finding such mutations, it offers a strategy that is cost effective - it is far more affordable to genotype a large number of samples than it is to exome capture and sequence each sample. The hypothesis is that rare mutations in immunological relevant exonic regions predispose to CD risk.

iii) **Determining genetic load of coeliac associated risk loci** - the aim is to compute a combined gene dose score of 58 current associated coeliac loci in coeliac individuals and controls to determine if affected individuals in families carry a higher proportion of GWAS-risk loci compared to healthy individuals in the same families and unrelated population cases and controls. The hypothesis is that affected individuals from coeliac families, where CD is highly clustered due to a genetic predisposition, carry more CD associated risk loci than controls.

4.3 Sample selection

66 coeliac cases and 119 related controls from 12 large multiply affected pedigrees were selected for linkage analysis (Table 4.1; see Figure 1, Appendix I-A for pedigree illustrations). Pedigrees DA, BRK, BRE, HMN, BD, BR, BUT, H and SDY were obtained from Professor Paul Ciclitira. These pedigrees have been previously used for family-based linkage studies in CD (Brett, Yiannakou et al. 1999; King, Yiannakou et al. 2000; King, Fraser et al. 2001; King, Moodie et al. 2002; King, Moodie et al. 2003).

For exome SNP case-control association analysis, the sample set consisted of 7,728 coeliac cases, 8,274 controls of European (UK) ancestry, with an additional 112 UK cases from coeliac pedigrees and 129 related controls used

for genetic load analysis (Table 1, Appendix I-A). The UK control set contained 5,430 UK 1958 Birth Cohort participants and 2,844 UK Blood Services-Common Controls. 7,728 affected coeliac individuals were diagnosed according to standard clinical criteria including Marsh Stage III small intestinal biopsy and compatible serology. Detailed case diagnoses are outlined in Trynka et al, online methods (Trynka, Hunt et al. 2011). 112 coeliacs from pedigrees were all diagnosed according to the revised ESPGHAN criteria (ESPGHAN 1990) .

Table 4.1: Linkage pedigrees

Family name	Sample size (Affected/Unaffected)	Affected Individuals (Genotyped/Total)	Unaffected Individuals (Genotyped/Total)	Number of exomes sequenced
DA	4/9	4/4	9/9	2
BRK	6/23	6/6	17/23	2
BRE	6/19	6/6	17/19	1
HMN	5/15	5/5	9/15	1
BD	6/10	2/6	4/10	1
BR	4/22	3/4	14/22	1
BUT	7/23	6/7	17/23	1
H	4/8	4/4	5/8	1
SDY	13/21	10/13	21/21	3
FAM008	8/6	7/8	2/6	1
FAM063	8/7	7/8	0/7	1
FAM014	6/22	6/6	4/22	2
Total	**77/192**	**66/77**	**119/185**	**17**

4.4 Experimental design and laboratory method

All samples were genotyped on the Infinium HD Immunochip custom array designed by Illumina. Genotyping was performed at Barts and the London Genome Centre and arrays were scanned on the Illumina iScan at the Institute of Child Health, University College London. NCBI build 36 (hg18) mapping was used. Chapter 2, section 2.6 details genotyping methods, as per Illumina's protocols.

Non-parametric linkage analysis was performed using all Immunochip SNP markers. Rare variants (0.5% frequency based on 1000G) from the exome

sequencing dataset were extracted from regions under linkage peaks ($p<0.01$) in all families.

Exome SNPs were selected for the Immunochip final design from both phase and two exome sequencing experiments: 1,526 exome SNPs from 60 young onset coeliac case samples captured by NimbleGen™ exome microarray and sequenced with Illumina GAII*x* (76bp paired end multiplex sequencing) and 1,336 SNPs highlighting interesting immune genes from four high coverage exome resequenced samples (multiply affected (>4) individuals per family), captured by NimbleGen™ in-solution capture and sequenced with Illumina GAII*x* 76bp paired end sequencing. Chosen exome variants were missense mutations in immune genes and all nonsense and frameshift mutations with an allele frequency difference with 1000G 2010 release CEU (Northern European ancestry) data. The final exome SNP Immunochip dataset consisted of 2,862 variants for genotyping.

From the most current coeliac fine mapping results, all SNPs reaching a $p=10^{-8}$ were extracted from the final immunochip dataset, totaling 58 variants including the HLA high-risk variant rs2187668. This dataset was used to test genetic load of disease associated risk variants in coeliac individuals.

4.5 Results: Linkage analysis with Immunochip SNP markers

4.5.1 Sample and data quality control

All analysis was performed with PLINK v1.07 (Purcell, Neale et al. 2007). Polymorphic SNPs were selected by performing a Mendelian check and removing individuals with an excess of mismatches i.e. those SNPs not inherited by father or mother due to genotyping errors or erroneous assignment of relative status in the pedigrees. Prior to this, all X-linked and CNV tagging SNPs were removed from the dataset. LD can inflate multipoint linkage analysis results so LD pruning was performed on founders only to remove any bias. A Hardy Weinberg equilibrium filter (HWE) of 0.001 and MAF of 0.2 was applied

and SNPs were pruned using an r^2 threshold of 0.2, leaving 3,700 markers for linkage analysis.

One individual each from SDY and BRK family was removed due to excess Mendelian errors. Clustering was performed based on IBS linkage in all individuals to determine if familial samples clustered with control HapMap3 samples. All family samples were merged with HapMap3 CEPH, CHB, JPT and YRU samples and pairwise identical-by-sharing was calculated resulting in overlapping clusters with HapMap3 CEU samples. There was one major outlier in the BRE family indicating mixed ethnicity.

4.5.2 Non-parametric linkage analysis

Subjects were classified as affected, unaffected or unknown affection status according to pedigree records obtained by Professor Paul Ciclitira, Dr Susan Neuhausen and us. Multipoint NPL analysis was performed with Merlin linkage analysis software (Abecasis, Cherny et al. 2002) based on the Kong and Cox LOD score statistic comparing alleles shared IBD for all affected individuals (Kong and Cox 1997). To prove the hypothesis that rare variants with large effect size result in large increases in allele sharing in families compared to common variants of small effect size, an exponential model was selected. Under the NPL null hypothesis that a locus is not linked to a susceptibility gene, the statistical behaviour of the number of IBD alleles depends on their relationship to each other, as determined by pedigree structure, and not on their disease status. For a linked locus to contain a disease-susceptibility gene there is an expected increase in the number of IBD alleles among affected individuals, relative to null expectation. The NPL statistic, or *P* value, reflects IBD alleles shared evaluating non-random segregation at chromosomal locations, which is reported here. Furthermore the LOD scores reported here, proposed by Kong and Cox, is maximized on a single parameter δ in the numerator based on the observed genotypes, representing the degree of allele sharing amongst affected individuals. Under the null, $\delta = 0$ and the alternative $\delta > 0$ corresponds to a high degree of allele sharing (Nyholt 2000). The score makes confidence comparisons

amongst loci asymptotically and is an extension of the NPL statistic (Kruglyak, Daly et al. 1996). The LOD score differs in a parametric test because it is based on the likelihoods of obtaining linkage given a set of assumptions in the underlying model.

Because the interpretation of linkage results should be performed in terms of the appropriate significance threshold on the pedigrees studied rather than the observed statistics, maximum *P* values corresponding to the power of core pedigrees to detect linkage was measured by running in-silico simulations assuming no locus heterogeneity. According to these simulations, BRE, BD, SDY, FAM008, FAM063 and FAM014 showed sufficient power to produce linkage *P* values of 0.0001 or more, determined by their maximum linkage *P* scores. A *P* of 0.0001 is equivalent to a linkage LOD score of 3. Table 4.2 shows maximum linkage *P* values and size of linkage regions at $P < 0.01$ Mb. All linkage graphs are in Appendix II.

Table 4.2: Summary of non-parametric linkage results

Family name	Maximum linkage *p*	Max observed linkage p	Size of linkage region at p < 0.01 (Mb)	Number of rare, LoF and non-synonymous variants in linkage region
DA	0.004	0.005	25.47	6
BRK	0.0013	0.0011	40.08	3
BRE	0.0004	0.0004	23.76	1
HMN	0.005	0.005	71.63	3
BD	0.0005	0.05	0	0
BR	0.002	0.004	46.04	0
BUT	0.0013	0.003	34.34	0
SDY	0.00002	0.03	0	0
FAM008	0.00001	0.0002	29.01	0
FAM063	0.0002	0.0008	52.42	2
FAM014	0.0002	0.005	26.98	3
H	0.02	0	0	0

Of the six families that had power of producing a linkage *P* value of 10^{-4}, only three had reached maximum observed linkage *P* values of 10^{-4} across all chromosomes: BRE, FAM008 and FAM063. For the 3700 markers genotyped, there was no evidence of linkage ($P<0.01$) between the marker SNPs or any disease locus on any chromosome for families BD, SDY and H.

The highest maximum multipoint NPL LOD score was observed in BRE (NPL LOD 2.40) at chromosomal position 5q33.3. The 5q region has previously been implicated in 110 Italian sib pairs and an independent set of Italian coeliac families (Greco, Corazza et al. 1998; Greco, Babron et al. 2001), but no strength of evidence was replicated in Finnish families (Liu, Juo et al. 2002). However a meta analysis of multiple European CD data established 5q31-33 (CELIAC 2) as a significant linkage region (Babron, Nilsson et al. 2003) highlighting increased power effects when using meta datasets. A multipoint NPL LOD score of 1.55 was observed for BUT at 5q35.3, but no other families provided evidence of linkage around this region.

BRK, BRE and FAM0063 were the only families where LODs of >2 were observed at chromosomes 2, 4, 5, 6, 7, 9, 11 and 19. The maximum observed linkage *P* values for these families were 0.001, 0.0004 and 0.0008, respectively. For BRE, the most significant LOD of 2.41, outside the HLA, was to 19p13.3; linkage to 19p13.1 has previously been implicated in a Dutch sib-pair family (Van Belzen, Meijer et al. 2003). Significant linkage on chromosome 11p11 was identified in 50 UK coeliac families but here linkage in BRK and FAM063 was identified at 11p15.5 (King, Fraser et al. 2001). Lower multipoint NPL LOD scores in the CELIAC1 locus on chromosome 6p21.32, containing HLA class II molecules, were observed in four out of the 12 families: 2.5, 1.2 and 1.71 and 1.5 in BRE, BUT and FAM008 and HMN respectively.

Families with a low HLA risk were prioritised as potentially harbouring rare non-HLA high-risk disease mutations. Figure 4.2 illustrates the number of different HLA genotypes in familial cases and controls. 46% of affected individuals were HLA-DQ2.5 homozygote or heterozygote, and 100% had at least one copy of either DQ2.5 or DQ8. No low risk HLA genotypes were observed in coeliac cases. Overall 74.5% (70/94) of unaffected individuals carried either HLA-DQ2.5 or

HLA-DQ8 homozygous or heterozygous genotypes, which is larger than the overall population (30% of general population carry DQ2.5/DQ8 molecules). Furthermore, of the 74.5% DQ2.5/DQ8 genotypes from unaffected individuals, 18.6% were from individuals who married into the family (13/70).

This information was of use to note why no linkage in the HLA region was observed in 8/12 pedigrees. Firstly, it is important to understand what is required for a family to be informative for linkage. For any linkage to be observed, alleles have to be transmitted IBD on the shared chromosome in an NPL model. Figure 4.3 illustrates the difference between alleles shared IBD and IBS. In this scenario, when both siblings have a genotype of 1,4, both type 1 and type 4 alleles are IBD and IBS. When the second sibling has a genotype of 2,1, no IBD is present because one type 1 allele is inherited from the mother and one is inherited from the father, so the ancestral source cannot be determined.

Figure 4.2: Graph of number of different HLA genotypes in affected and unaffected individuals from 12 linkage families

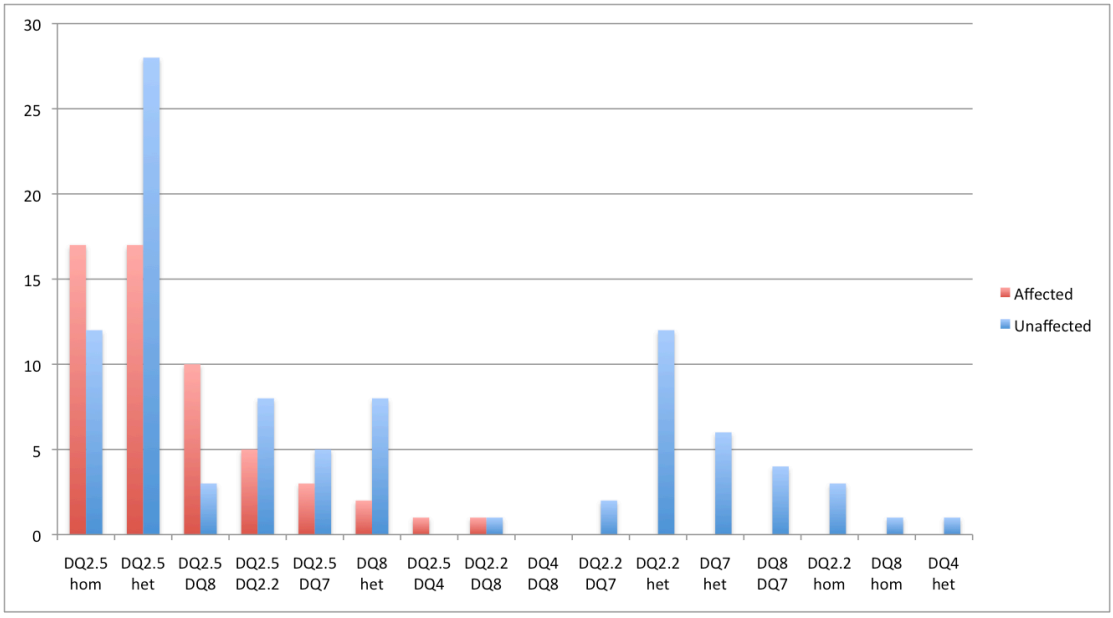
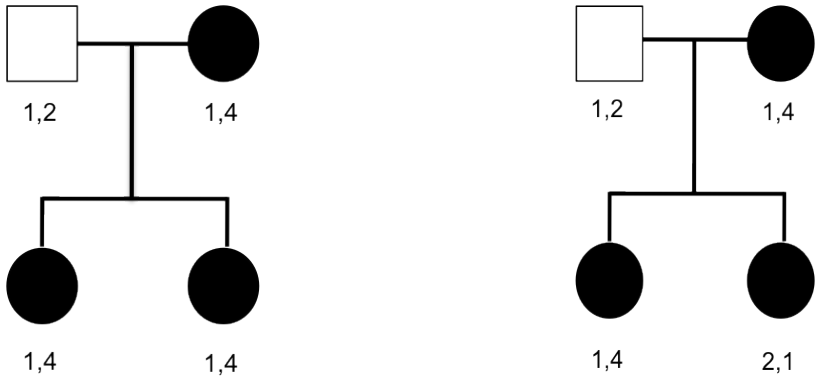


Figure 4.3: Alleles shared IBD and IBS in sibling pairs when allele sharing differs in second sibling, assuming the marker is unlinked to the disease



Alleles shared IBD: 1 and 4
Alleles shared IBS: 1 and 4

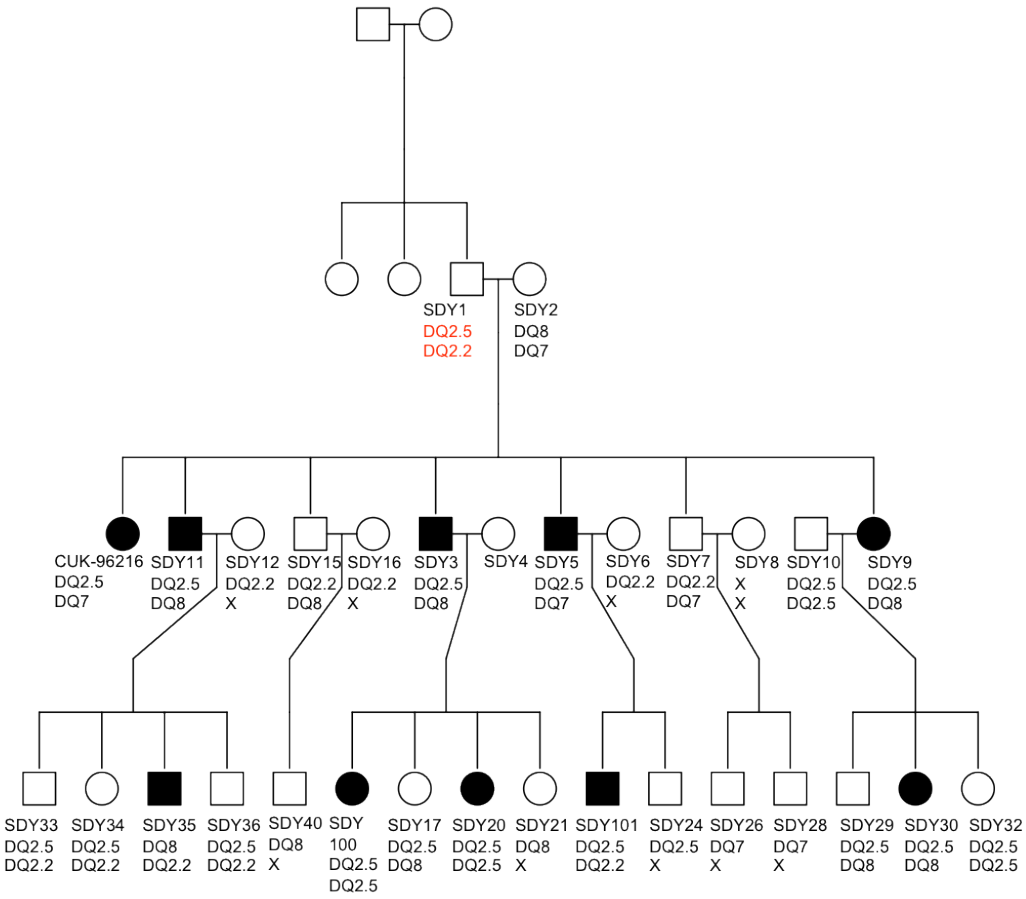
Allele shared IBD: no IBD present
Allele shared IBS: 1

Adapted from Teare & Barrett (Teare and Barrett 2005). First number is paternally inherited allele; second number is maternally inherited allele. No inbreeding is allowed for this situation to hold true.

This example was extrapolated in all linkage pedigrees to concur why they were not informative for linkage, using the HLA class II allele genotypes. As an example, SDY family is illustrated in figure 4.4 and all HLA genotypes for 11 remaining families are illustrated on the pedigrees in Figure 1, Appendix I-A. For SDY a proportion of unaffected individuals also carry the DQ2.5 risk haplotype so it is hard to infer where the coeliac risk allele comes from. An unaffected subject SDY9 carries the DQ2.5 allele, shared with 5 affected members of the family. SDY10, the unaffected husband of SDY9, carries two copies of the DQ2.5 risk allele. Their offspring SDY29 and SDY30 both carry one copy of the DQ2.5 allele, which they have inherited from their unaffected father. So the pedigree cannot be counted as informative because, given that SDY10 is homozygous for DQ2.5 and SDY30 and SDY29 carry only one copy of the allele, they can only have inherited them by this route. They both carry DQ8 on the other chromosome, inherited from the affected mother. The same is true for SDY100 and SDY20; since they are both homozygous for DQ2.5, one has

to have been transmitted from the unaffected mother. This observation proves transmission of risk alleles IBS rather than IBD, and hence why no linkage was observed at the HLA region. Brett et al. (1999) also reported the same results for family SDY, BRK AND H for HLA linkage (Brett, Yiannakou et al. 1999).

Figure 4.4: HLA genotypes IBD and IBS for SDY family



Red parental genotypes inferred from offspring genotypes. 'X' denotes other genotype. Third generation affected siblings all acquired DQ2.5 risk allele from their father, SDY1. Fourth generation transmissions are mostly IBS: DQ2.5 alleles for SDY100, SDY17, SDY20 and SDY21 cannot come from the same haplotype as the father, due to offspring homozygous genotypes, so these alleles are IBS. The same is true for SDY29, SDY30 and SDY32.

4.5.3 Analysis of exome variants in linkage peaks

To search for rare coding variants in linkage regions with a *P*>0.01, the following filters were applied to individual exomes from linkage families: i) MAF <0.5%, ii) only LoF variants, iii) regions without duplications. In total 18 rare nonsynonymous variants from individual familial exome sequencing datasets were identified in linkage regions, summarised in table 4.3. No rare indels were observed under linkage peaks. To test whether the identified SNPs were present on the same haplotype in all affected members of the linkage pedigree, all SNPs were Sanger sequenced in every affected member of the family. Ten out of 18 SNPs were validated in all affected individuals from five pedigrees, so were present on the same ancestral haplotype. The 10 genes harbouring these validated SNPs were selected for candidate gene resequencing, together with 17 genes from Chapter 3.

Table 4.3: Non-synonymous SNPs located in linkage regions ($p < 0.01$)

Family	Gene	Function	Chr: position	SNP	PolyPhen Prediction	dbSNP132 ID/function	Cases validated/ Cases tested
BRK	*FAM179A*	nsSNP	2:29259543	c.2555T>C	-	rs72788155/ missense	2/6
BRK	*NLRC4*	nsSNP	2:32474767	c.2166T>G	Probably damaging	-	6/6
FAM063	*EPAS1*	nsSNP	2:46607609	c.1798G>A	Possibly damaging	-	7/7
FAM063	*STON1*	nsSNP	2:48809609	c.1837C>G	Probably damaging	-	2/7
DA	*ARHGAP25*	nsSNP	2:69040504	c.739G>A	Probably damaging	rs61758703/ missense	4/4
FAM014	*IQGAP2*	nsSNP	5:75969341	c.3136G>T	-	-	1/6
FAM014	*DMGDH*	nsSNP	5:78293933	c.2573A>C	Probably damaging	-	4/6
HMN	*KIF13A*	nsSNP	6:17826085	c.1700A>C	-	-	5/5
BRE	*BRD2*	nsSNP	6:32942277	c.68G>A	Probably damaging	rs55650502/ missense	4/6
HMN	*GRM4*	nsSNP	6:34101193	c.81G>A	Benign	-	5/5
HMN	*TULP1*	nsSNP	6:35471412	c.1247G>A	Probably damaging	-	5/5
BRK	*SYTL2*	nsSNP	11:85445365	c.1004C>G	Probably damaging	rs74718633/ missense	2/6
DA	*ABCA9*	nsSNP	17:67039672	c.758C>T	Possibly damaging	-	4/4
DA	*KCNJ16*	nsSNP	17:68129412	c.1184A>G	Benign	-	4/4

DA	*SDK2*	nsSNP	17:71431712	c.1072C>T	-	-	1/4
FAM014	*MALT1*	nsSNP	18:56402558	c.1567G>A	Probably damaging	-	6/6
DA	*ACOT8*	nsSNP	20:44470575	c.862C>T	Probably damaging	-	4/4
DA	*EYA2*	nsSNP	20:45808514	c.1267C>T	Possibly damaging	-	1/4

Chromosome positions correspond to the same linkage regions in each family

4.6 Results: Exome SNP case control association

A Fisher's exact test was conducted to generate a *P* value of significance comparing exome SNPs between 7,728 UK coeliac cases and 8,274 controls. Prior to any analysis, very low call rate variants were removed. Illumina GenomeStudio GenTrain2.0 algorithm was used to cluster all samples, and clusters were manually re-adjusted or excluded for variants with low quality statistics (call rate <99.5%, low GenCall score, and high-intensity no-calls). This left 1,932 out of 2,862 variants passing quality control for the final association dataset. The 7,728 coeliac case and 8,274 control dataset had previously undergone quality control steps including exclusion for call rate <99.5%, incompatible gender, duplicates, first or second degree relatives and ethnic outliers (identified by multi-dimensional scaling plots of samples merged with HapMap3 data) (Trynka, Hunt et al. 2011).

In the final association analysis, SNPs with MAF >5% and known HLA SNPs were removed, leaving seven SNPs from five loci that showed association at $P<10^{-5}$ (Table 4.4). Odds ratios ranged from 0.848 to 1.583 for all seven associations. Five SNPs ($P<10^{-5}$ - $P<10^{-7}$) did not reside in any genes exhibiting good functional candidacy for CD. SNP imm_15_77018533 is in a non-coding region of *SCAPER*; this gene is only expressed in the pancreas. Two SNPs on the X chromosome (vh_X_5821532 and vh_x_5831768) in *NLGN4X*, belonging to a family of neuronal cell surface proteins, and imm_3_46561626 in *LLRC2* also does not have any type of immune mediated disease function. SNP rs1800562 in the *HFE* gene is just outside of the HLA region on 6p22.2, similar to two associations reaching $P<5 \times 10\text{-}^{8}$: vh_6_24672519 ($P=2 \times 10^{-14}$ OR = 1.6) and vh_6_24684610 ($P=1.8\text{x}10^{-13}$ OR = 1.5) (Figure 4.5). A conditional logistic regression was performed on these SNPs to test for independent associations.

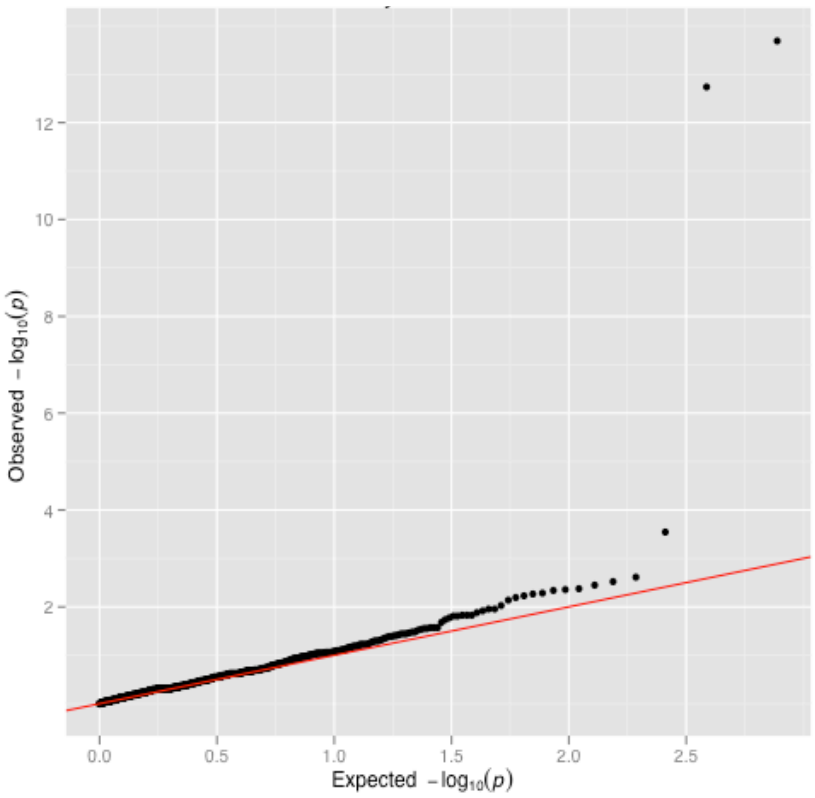
Table 4.4: Fisher Exact test results for rare exome SNPs in coeliac UK dataset at p < 0.01

Chr	SNP	Position	Allele 1	F_A*	F_U**	Allele 2	*P* value	OR
1	imm_1_159289698	159289698	C	0.007246	0.004774	T	0.004532	1.522
2	imm_2_62081517	62081517	T	0.02206	0.01777	C	0.006349	1.247
2	vh_2_233605747	233605747	A	0.003947	0.006164	G	0.005832	0.6389
3	**imm_3_46561626**	46561626	C	0.09181	0.07638	T	6.79E-07	1.222
4	imm_4_103165415	103165415	A	0.06515	0.05584	G	0.0005054	1.178
6	**vh_6_24672519**	24672519	G	0.04497	0.02889	A	2.04E-14	1.583
6	**vh_6_24684610**	24684610	T	0.04587	0.03015	C	1.82E-13	1.546
6	**rs1800562**	26201120	A	0.06308	0.07766	G	3.58E-07	0.7996
6	imm_6_83723885	83723885	G	0.01928	0.02387	A	0.005005	0.804
6	vh_6_84164919	84164919	A	0.01928	0.02532	T	0.0002803	0.7567
7	imm_7_22737681	22737681	T	0.0187	0.02302	C	0.007012	0.8085
8	vh_8_38184403	38184403	G	0.004076	0.006406	C	0.004221	0.6348
9	imm_9_35368401	35368401	T	0.01579	0.01994	C	0.00538	0.7883
10	imm_10_101995610	101995610	T	0.01766	0.01402	C	0.00922	1.265
10	imm_10_112714563	112714563	T	0.01792	0.02266	C	0.002957	0.787
11	vh_11_56267736	56267736	G	0.08534	0.09439	A	0.004886	0.8951
11	imm_11_66374366	66374366	A	0.01792	0.01384	G	0.003545	1.3
11	imm_11_127837472	127837472	A	0.01754	0.0223	G	0.002383	0.7827
15	**imm_15_77018533**	77018533	C	0.1455	0.1302	T	7.41E-05	1.137
19	vh_19_7114065	7114065	A	0.07078	0.07856	G	0.008376	0.8935
23	imm_X_140860	140860	T	0.2325	0.2158	C	0.002175	1.101
23	imm_X_219362	219362	A	0.1469	0.1328	G	0.001624	1.124
23	imm_X_1388421	1388421	T	0.05593	0.04828	G	0.006455	1.168
23	imm_X_1427404	1427404	C	0.09083	0.07947	A	0.001425	1.157
23	imm_X_1457644	1457644	C	0.1618	0.149	G	0.006401	1.103

23	**vh_X_5821532**	5821532	A	0.1402	0.1613	G	2.87E-05	0.848
23	**vh_X_5831786**	5831786	A	0.1068	0.0892	G	3.03E-06	1.221

SNPs in HLA region were removed *Minor allele frequency in cases ** Minor allele frequency in controls (as reported in PLINK v1.07)
OR = odds ratio

Figure 4.5: Q-Q plot of Fisher Exact test *P* values



SNPs with MAF >5% in known disease loci were removed. Two SNPs deviating from expected *p* values are vh_6_24672519 (P=2 x 10^{-14} OR = 1.6) and vh_6_24684610 (P=1.8x10^{-13} OR = 1.5).

4.6.1 Conditional logistic regression

In the UK only dataset, the r^2 and D' values for vh_6_24672519 were 0.013 and 0.400, and for vh_6_24684610 the values were 0.013 and 0.413, respectively, suggesting independent effects of both SNPs on chromosome 6. Due to high LD in the HLA region on chromosome 6 (close to where both SNPs are located), a conditional logistic regression was performed to test for independent effects. The regression analysis result showed that the HLA-DQ2.5 SNP rs2187668, which has the largest effect in CD, accounts for all significance, so neither SNP is independently associated with CD risk (Table 4.5).

Table 4.5: Results for conditional logistic regression for two associated SNPs, vh_6_24672519 and vh_6_24684610

HLA SNP	r^2	D'	Associated SNP	OR	*P* value
rs2395182	0.001	0.350	vh_6_24672519	1.527	1.29E-11
	0.001	0.390	vh_6_24684610	1.484	1.42E-10
rs2187668	0.011	0.404	vh_6_24672519	0.9352	0.3758
	0.011	0.387	vh_6_24684610	0.9248	0.2942
rs7775228	0.002	0.600	vh_6_24672519	1.64	1.46E-15
	0.002	0.548	vh_6_24684610	1.597	1.55E-14
rs4713586	0.000	0.617	vh_6_24672519	1.585	1.01E-13
	0.000	0.579	vh_6_24684610	1.55	6.58E-13
rs7454108	0.000	0.178	vh_6_24672519	1.581	1.52E-13
	0.000	0.148	vh_6_24684610	1.546	9.24E-13

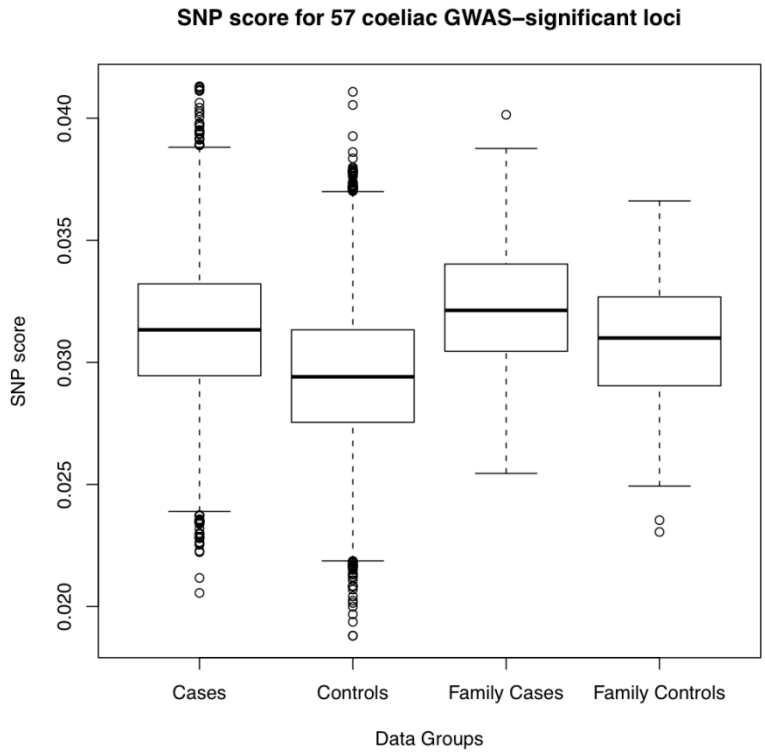
OR = odds ratio**.** **r^2** and D' are measures of LD; complete and perfect LD is when D' = 1 and r^2 = 1 i.e. two SNPs are co-inherited 100% of the time, respective of allele frequencies, e.g. r^2 will only be 1 if there are 2/4 possible haplotypes but allele frequencies are the same**.**

4.7 Results: Current coeliac associated loci contribution in coeliac individuals

57 independent non-HLA SNPs associated with CD risk were subjected to SNP score analysis using a SNP scoring algorithm in PLINK (v1.07). The score itself was calculated based on the logarithm of allelic odds ratios to identify whether certain individuals contained more of a set of associated variants. This provided a quantitative measure of genetic load in coeliac cases compared to controls. The analysis compared all UK coeliacs (7,728), all UK controls (8,274), all coeliacs from large pedigrees (112) and pedigree related controls (129). Further analysis compared SNP scores between groups of cases stratified against the number of affected coeliacs per family. This analysis used UK coeliac cases for which family information was available plus the 112 coeliacs from large pedigrees.

Box and whisker plots were produced in R (version 2.13) illustrating SNP scores for 57 coeliac associated loci and 58 coeliac associated loci including the HLA-DQ2.5 SNP rs2187668 (Figure 4.6). A Wilcoxon rank test between all cases and controls, assuming the data are independent and come from the same population, highlighted a significant difference in SNP score across all loci (Wilcoxon $p<2 \times 10^{-16}$). The SNP score was much closer to the median for all groups without rs2167668, but the *P* value remained the same (Wilcoxon $p<2 \times 10^{-16}$). A modest positive correlation in box plot medians between the number of coeliac individuals per family and SNP score was observed, more so with the inclusion of rs2187668 (Figure 4.7). A significant Kruskal-Wallis test *P* value of 1.507^{-11} confirmed this observation. Here, a non-parametric test assumes independent groups, hence was chosen over a parametric test i.e. an Anova. Analysis with 57 SNPs only reduces the Kruskal-Wallis p value to 3.03^{-05} and SNP scores have a larger distribution around the median. Interestingly, the SNP score for HLA-DQ2.5 SNP rs2187668 on its own is increased when there are more than four affected individuals in a family (Figure 4.8).

Figure 4.6: SNP score for 57 coeliac risk loci, with and without HLA SNP rs2187668



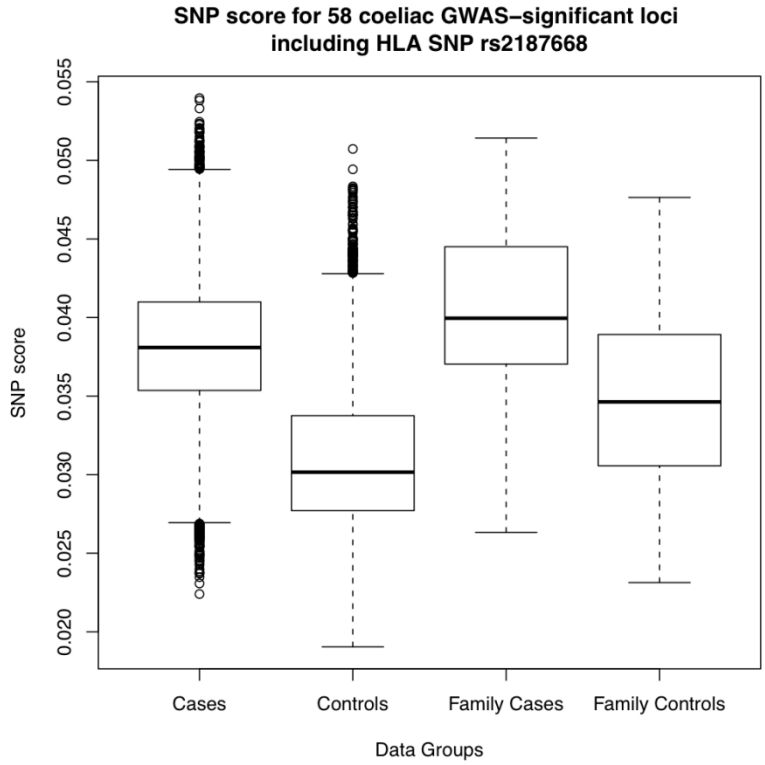


Figure 4.7: SNP score for 57 (58 with HLA SNP rs2187668) coeliac risk loci stratified against number of affected individuals per family

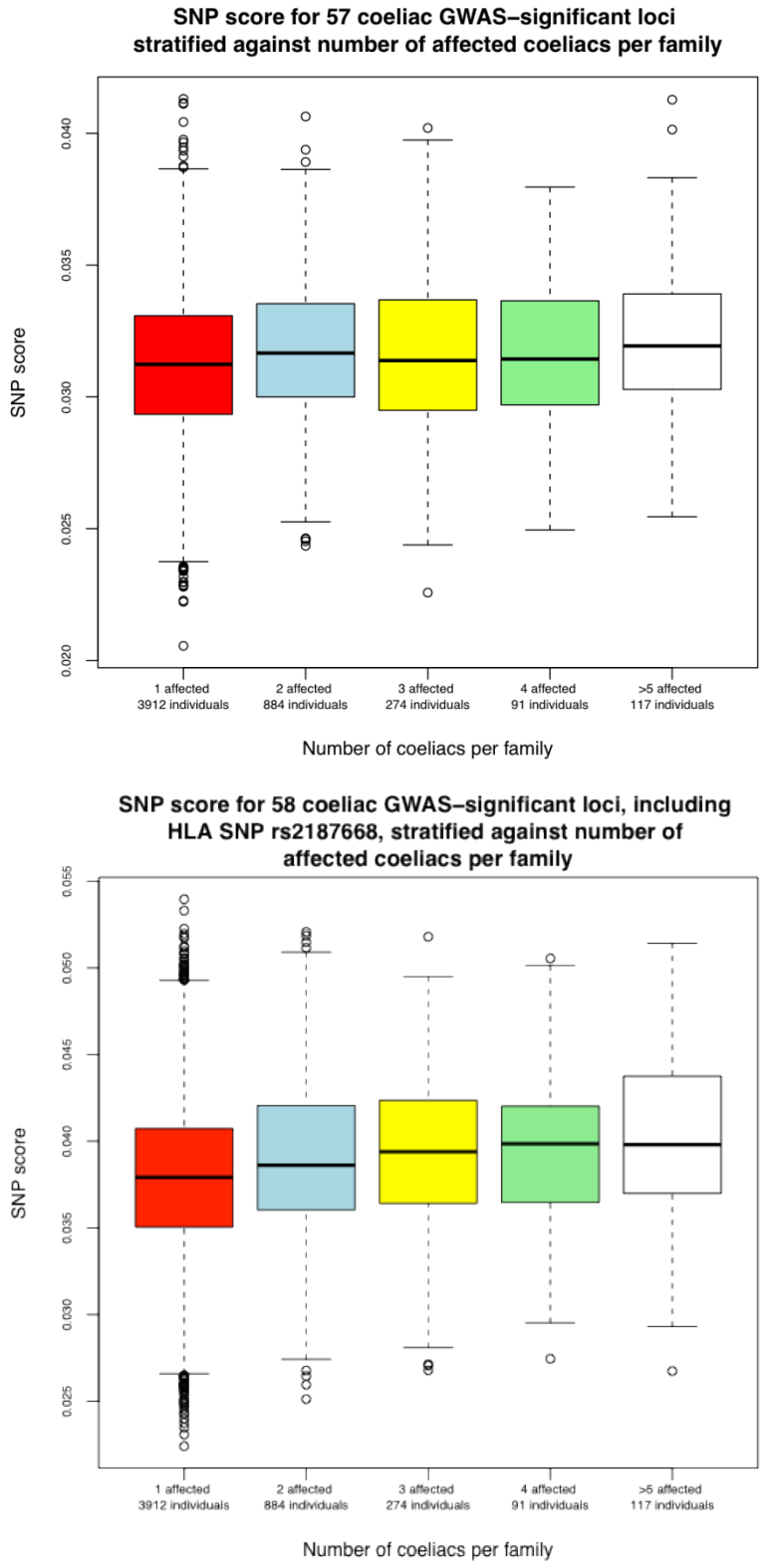
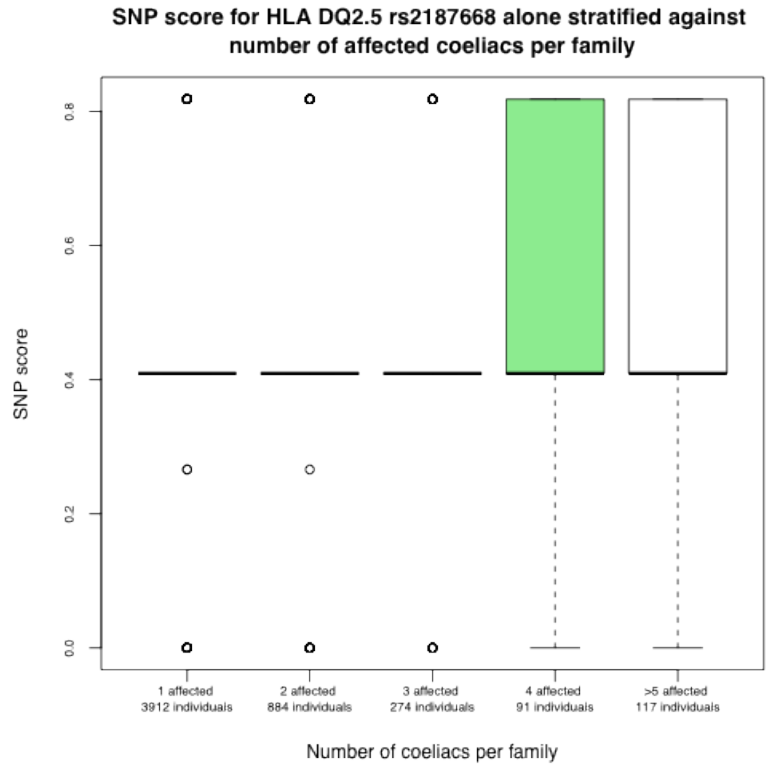


Figure 4.8: SNP score for HLA DQ2.5 rs2187668 alone vs. number of affected individuals per family



4.8 Chapter Discussion

The content in this chapter used 3,700 post quality control markers from the Immunochip dataset to compute NPL linkage analysis in 12 coeliac pedigrees, 1932 post quality control exonic SNPs for a case control association test and 58 coeliac associated SNPs for gene dose analysis.

The hopeful outcome of any case control association study is to find highly associated SNPs in disease cases when compared to matched controls. The decision to put exonic SNPs on the Immunochip array was to give early insight of any rare causal mutations by testing them in thousands of cases and controls. A normal GWAS array has polymorphic SNPs spanning the entire genome; Immunochip contained custom content of which contained 1,932 SNPs (post quality control) from very early exome data (aligned to reference human genome build 36) out of 196,524 variant markers (passing quality control at

Illumina). A Fisher exact test was applied to search all loci, with exonic SNPs P_{GWAS} between $p<10^{-5}$ and $p<10^{-8}$, for genes with good functional candidacy for CD. No exome SNPs with roles in CD, or any overlapping autoimmune disease (Chapter 1, Table 1.1) reached significance between $p<10^{-5}$ and $p<10^{-8}$ in the Immunochip dataset in 7,728 coeliac cases and 8,274 UK controls. The top SNPs, vh_6_24672519 and vh_6_24684610 mapped to a non-coding region of *ACOT13* on chromosome 6p22.3, just outside the MHC. Conditional logistic regression confirmed non-independent effects of both SNPs from the strongest HLA DQ2.5 SNP, rs2187668. Significantly, these SNPs were chosen from very early pilot data; some early analysis flagging 139 exome SNPs present on the Immunochip found 43% to be false positive, and approximately a third of possible risk variants were not genotyped after stringent quality control. It is fair to conclude that due to this high false positive rate no real association was found in the exome dataset. In spite of no significant association result, the Immunochip dataset proved invaluable in confirming coeliac pedigree relationships. Pedigree structures were confirmed by applying PLINK segmental sharing methods. A simple pairwise IBD estimation using SNP by SNP identity by state sharing has a sensitivity to detect only down to 1st cousin relationships (pi-hat=12.5% sharing, 3 meioses separation) but PLINK segmental sharing can identify individuals separated by 6 meioses or more. This validated relationships and those with disconcordant pi-hat values according to pedigree structure were removed.

Combining linkage analysis with exome sequencing data was a strategy chosen to pinpoint causal functional variants in regions where excess allele sharing was evident; model free linkage was applied to identify IBD alleles and any ambiguous evidence of linkage containing false positive loci was excluded as only variants under the linkage peak were analysed. Linkage analysis in families needs to be highly powered in order for significant LOD scores to be produced and this was addressed by using multigenerational families instead of sib-pairs due to evidence of the latter being underpowered in previous coeliac studies (Clot, Fulchignoni-Lataud et al. 1999; Greco, Babron et al. 2001). One would expect that the disease trait would have direct correspondence with the

underlying alleles for the genotyped SNPs, as they were chosen from immuno-biologically relevant loci, and if disease risk were due to a rare variant then one would expect to see that affected individuals would tend to inherit the rare risk alleles from the same ancestral source. Ng et al. importantly pointed out in one of the earliest exome sequencing studies that the large number of private mutations in a single exome is a caveat when trying to identify causal variant(s). They applied intersection filtering in individual exomes with a monogenic disorder, Freeman Sheldon Syndrome, to identify a single gene, *MYH3*, containing variants in multiple cases (Ng, Turner et al. 2009). Their most important finding was rare causal nonsynonymous variants were shared on the same haplotype amongst affected individuals, proving that an exome sequencing study can serve as a genome wide scan. The linkage method here was to identify those shared haplotypes possibly harbouring exome variants.

The distribution of HLA genotypes in CD cases and controls made the grandparental origins of alleles unclear, hence evidence of linkage was only observed in three families. For the majority of pedigrees, coeliac-associated common HLA alleles, required for the HLA-DQ2.5 and DQ8 genotypes, from members marrying into the family resulted in uninformative pedigrees for linkage, by way of being IBS rather than IBD. So, although the power issue can be explained in terms of HLA linkage, it did not deter from using the same pedigrees for the purpose of the aims in this chapter. Furthermore, replication of HLA linkage has been unsuccessful in other coeliac linkage studies due to a lack of distinction between alleles IBD and IBS, low marker density and no differences in inheritance patterns between affected and unaffected members (Eller, Vardi et al. 2006; Vidal, Borg et al. 2009). In support of the LOD scores obtained here, previous linkage studies using families provided by Professor Paul Ciclitira reported maximum LOD scores of 1.9 at 10q23.1 and 16q23.3 and 1.5 at 11p11 (King, Yiannakou et al. 2000), which increased to 2.6 when an additional 34 families were included, using microsatellite markers, which remain highly informative measures of associating linkage with disease due to their greater allelic and haplotype diversity. Nowadays, a SNP based scan is a more practical approach, but here linkage was attempted across a set of GWAS-

risk associated markers and essentially the 186 loci on the Immunochip array did not offer full coverage of the genome. Given that, suggestive linkage (P=10^{-4}) was detected in BRE, BRK and FAM0063 at, or close to, previously implicated non-HLA CD linkage regions 5q33.3, 19p31.1 and 11p11, respectively. Importantly, the linkage method applied here was another strategy to decrease the abundance of variants in the exome sequencing data, however having larger, extensive families and more exomes to search in would have provided the power required for a more definitive result. A recent study by David Kelsell (personal communication at the Blizard Institute) identified that 55% of 44 exomes from affected atopic eczema families have mutations in *FLG*, a gene largely associated with this phenotype. This result was obtained by whole exome sequencing of all members of the pedigrees, and a highly significant *P* value, comparing *FLG* mutations in cases and controls (6×10^{-11}), supports the advantage of an extensive exome dataset.

Excess HLA risk allele types identified in pedigrees was supported by genetic load analysis in affected individuals. An increase in SNP scores was evident in familial cases compared to all population controls, especially as the number of affected individuals per family increased. However, the HLA-DQ2.5 rs2167668 SNP was probably the driving force of this increase, with the remaining 57 non-HLA loci having low additive scores overall, as highlighted when HLA-DQ2.5 was analysed alone. This substantiates known gene dosage effects of HLA alleles in CD individuals (Murray, Moore et al. 2007) particularly the DQ2.5 gene, which has the largest odds ratio (OR=8) compared to 57 GWAS risk loci (highest to lowest OR's = 1.70-0.71).

Key to the aim of applying linkage was to recognize that only a few families in a sample contribute significantly to a linkage signal, so a search for mutations can be targeted to a small number of families. With that in mind, the analysis was successful in identifying shared chromosomal regions, and to complete the candidate gene list, exome variants under linkage peaks reaching LOD scores >1 were assessed. The SNVs in candidate have shown to be inherited IBD in all affected members of the pedigrees by the linkage test and are present on the same haplotype, as confirmed by Sanger sequencing. Gene selection for

targeted resequencing was based on, in order of priority, validation in all affected individuals, cDNA size and known immune function (Table 4.6).

Table 4.6: Candidate genes from linkage analysis selected for targeted gene resequencing

Gene	cDNA size (bp)	No. of exons	Known immune?	Validated true positive in all cases?
NLRC4	3,581	10	No	Yes
EPAS1	5,160	16	Yes	Yes
ARHGAP25	2,979	11	No	Yes
GRM4	3,879	10	No	Yes
TULP1	2,162	15	Yes	Yes
KCNJ16	4,002	5	No	Yes
MALT1	8,789	17	Yes	Yes
ACOT8	1,168	6	No	Yes

4.9 Chapter Conclusions

The points below conclude the findings from the research in this chapter:

1. Linkage peaks with LOD scores >1 were observed across 12 coeliac pedigrees using NPL analysis.
2. No significant associations were found with 1,932 post quality filtered exonic SNPs selected from the exome sequencing dataset (phase one and two).
3. Excess HLA risk allele types were identified in 12 coeliac pedigrees and in individuals where >4 individuals in the family has the disease.
4. Eight genes from the linkage analysis, with variants shown to occur on the same haplotype in the tested family, have been taken forward for candidate gene resequencing (Chapter 5).

Chapter 5

Exome study candidate gene resequencing in 2,304 coeliac cases and 2,304 controls

5.1 Introduction

The content in this chapter describes simultaneous amplicon sequencing-based variant discovery and genotyping for coding exons in 24 exome-candidate genes in 2,304 UK coeliac cases and 2,304 matching controls. This final project is a follow-up study of candidate genes chosen from shared exome variant analysis in familial exomes (Chapter 3), variant segregation analysis from multigenerational families (Chapter 3), an aggregate gene-based test for rare variants (Chapter 3) and non-parametric linkage analysis (Chapter 4).

Recent large-scale human sequencing studies have revealed an abundance of rare variants (defined as MAF <0.5%) that are geographically localized and likely to have deleterious functional consequences compared to their common counterparts. A recent study resequenced 202 genes in 14,002 people and found 95% of coding variants to be rare of which 74% were observed in only one or two people (Nelson, Wegmann et al. 2012). This study and others showing similar results across ~15,000 genes demonstrate that most rare allelic mutations of any given sample will be unique and only detectable by direct resequencing of the sample (Tennessen, Bigham et al. 2012; Fu, O'Connor et al. 2013). Importantly, these latter studies have also discovered that excess of rare variants is due to recent population growth and large samples sizes are required in order to associate them with a complex phenotype, supporting findings from population genetic studies (Cargill, Altshuler et al. 1999; Williamson, Hernandez et al. 2005).

In order to discover rare variants contributing to disease and test them for association with phenotype the overall proposed method has been to resequence a small initial sample size and then genotype the discovered variants in a larger sample set (Nejentsev, Walker et al. 2009; Momozawa, Mni et al. 2011; Rivas, Beaudoin et al. 2011). Nejentsev et al and Momozawa et al only located low frequency (MAF 0.5-5%) coding mutations in *IFIH1* (MAF in controls 0.67-2.22%) and *IL23R* respectively, where the small sequencing sample size failed to determine a large fraction of the rare variants present. The best example describes a mutational analysis of *CARD15/NOD2* mutations in

453 Crohn's disease patients where 3/67 rare associated variants were more frequent in cases, but exhibited an allele frequency of >0.05 (Lesage, Zouali et al. 2002). A pooled sequencing study in RA targeted 25 GWAS associated candidate genes based on gene relationships across implicated loci (GRAIL) pathway analysis for exon resequencing but highlighted only relatively weak associations in *IL2RA* and *IL2RB* (missense variant burden signal of association p=0.007 and p=0.018, respectively) (Jordan, Cao et al. 2012; Diogo, Kurreeman et al. 2013). A study in idiopathic generalized epilepsy, using similar methods to ones described in this thesis, failed to identify any rare risk disease associations after genotyping 3,897 candidate variants (878 cases, 1,803 controls) from an exome sequencing dataset of 119 subjects with two forms of the disease, indicating the lack of statistical power in the genotyped dataset to detect a true association. The study highlighted that the variants were sufficiently rare that each one only accounted for a small fraction of individuals and sequencing might have provided a better resolution of these variants (Heinzen, Depondt et al. 2012). Collectively, these studies have shown that investigating risk alleles in protein-coding regions in associated loci can identify genes of biological relevance in complex traits, but testing the entire rare variant content in a large case control sample set is necessary to implicate large risk effects. For this project, gene sequencing was chosen in place of genotyping the variants found in them to allow assessment of the full exon sequences rather than just information on sites over the exons, limiting areas for rare variant searching. In addition, if there is a high false positive rate in exome target capture, genotyping exonic SNPs will not be anymore informative and hence less cost effective. Performing highly multiplexed sequencing of high quality and deep coverage will enable direct genotyping in a case-control sample set.

Many statistical methods have been proposed to detect rare variant associations in common diseases. The analysis of rare variants is complicated due to the fact that the power to detect a single rare SNV is dependent on its MAF. To overcome the issue of power in a scenario where tens of thousands of individuals cannot be sequenced, new methods have been developed that aggregate rare variants across a target region and incorporate, for example, if

the variant is a risk or protective variant and it's function and effect on the translated protein. Madsen and Browning proposed weighting variants based on their estimated frequencies in controls, where variants with a low frequency are given a higher weight compared to higher frequency variants (Madsen and Browning 2009). This differs from Li and Leal's method based on testing whether the rare variants present in cases are proportionately different compared to controls (Li and Leal 2008). Price et al extended the Madsen and Browning method by including the functional effect of the variant in the weighting scheme (Price, Kryukov et al. 2010). The common factor to all these tests, and others, is to group variants in a gene or candidate region and perform a gene-based test, instead of one test per variant per gene (Li and Leal 2008; Madsen and Browning 2009; Bhatia, Bansal et al. 2010; Han and Pan 2010; King, Rathouz et al. 2010; Liu and Leal 2010; Price, Kryukov et al. 2010; Ionita-Laza, Buxbaum et al. 2011; Neale, Rivas et al. 2011). The main advantages of these approaches are that statistical power is increased by: a) the burden of multiple testing is reduced as the number of regions containing aggregated SNVs is much lower than the number of overall SNVs; b) the combined allele frequencies of aggregated SNVs are higher than the individual allele frequency of each rare SNV. Gene-based tests based on these developments have been applied in the dataset described in the next sections.

5.2 Aim and hypothesis

The aim of the final follow-up study in this chapter is to determine if there are any rare (MAF <0.5%) functional variants in the coding regions of 24 candidate genes in coeliac cases or controls. These variants may be either protective or have a pathogenic risk to disease and will be found through PCR of target genes and highly multiplexed sequencing. All exons in candidate genes will be resequenced in a large sample size to elucidate the complete rare fraction of all coding regions. By treating sequence data like genotype data, single variant association and gene burden tests will be performed on all coding variants. The hypothesis is that rare variants might exist in candidate genes selected from

exome sequencing of multiply affected families (using a combination of linkage, shared variants analysis between multiple related subjects and gene burden tests for multiple potentially causal variants) that account for familial clustering in coeliac disease. The missing heritability for disease might lie in the rare coding mutation region of the allelic spectrum in candidate genes, supporting the CDRV hypothesis (Pritchard 2001; Bodmer and Bonilla 2008).

5.3 Pilot study

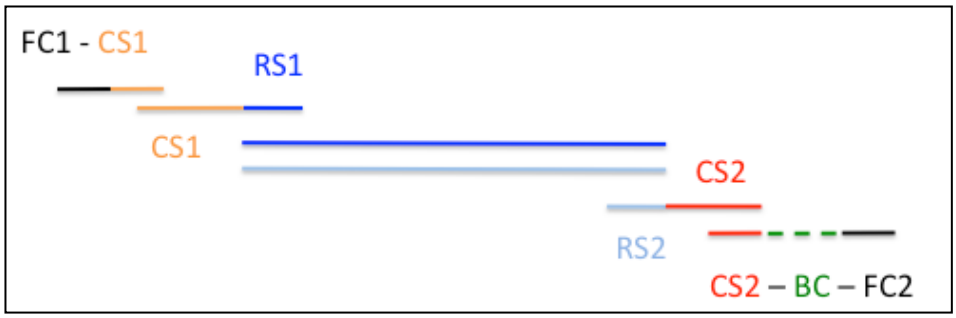
A pilot study was carried out using multiplex PCR technology designed by Fluidigm in order to test whether their multiplex amplicon tagging method could be implemented into a project containing large sample sizes. The exons of 48 genes from coeliac Immunochip significant regions were sequenced in 384 coeliac samples. The main aims of the pilot study were: to assess sequence data quality by testing the concordance rate of sequenced SNPs against Immunochip genotyped SNPs; to determine what sequencing read pair length provided the best data quality; to test coverage of reads and barcode evenness across samples; to assess read length distribution of aligned reads; to assess clonality of reads.

5.3.1 Fluidigm 48.48 Access Array™ Intergrated Fluidic Circuit technology

Fluidigm 48.48 Access Array™ Intergrated Fluidic Circuit (IFC) technology incorporates PCR into sample preparation for up to 480 target regions and 384 individual barcodes. One Access Array can multiplex PCR 480 primer sets (equating to 480 regions of interest) in 48 samples. The PCR takes place in the centre of the microfluidic array, incorporating all target regions plus 48 barcodes for each sample. Once the PCR products are harvested from eight separate arrays, they are pooled to create one high throughput-sequencing library containing 480 amplicons from 384 individuals. Each sample in the library contains flow-cell sequences and a barcode sequence on the 5’ end that

can be read using the standard Illumina indexing protocol (Figure 5.1). The clear advantage of this technology is that it allows multiple regions of interest to be multiplexed in one PCR reaction with several samples, whilst allowing high sequencing coverage necessary for rare variant analysis.

Figure 5.1: Primer set up for Fluidigm multiplex amplicon tagging for Illumina high throughput sequencing



Key: FC - Illumina flow-cell primers; BC – Fluidigm barcodes; CS – adapter sequences; RS – region specific primers with CS adapter sequences at 5' end

5.3.2 Pilot study Method

The coordinates for 48 exonic regions chosen from Immunochip loci were submitted to Fluidigm for a custom assay design and validation. Once received, primers and 384 coeliac samples were sent to Fluidigm's Research and Development site in Paris, France for processing on the Access Array IFC machines. The 48-plex library was sequenced on the Illumina GAII*x* at two sites in the UK: six 140bp, 10bp index, bidirectional single end runs were sequenced at Barts and the London Genome Centre and one 100bp, 10bp index, bidirectional single end run was sequenced at Cancer Research UK (CRUK), Cambridge. Bidirectional sequencing was selected as it allowed sequencing of the 5' and 3' ends in one read, cutting the cost of a paired-end read (the 384

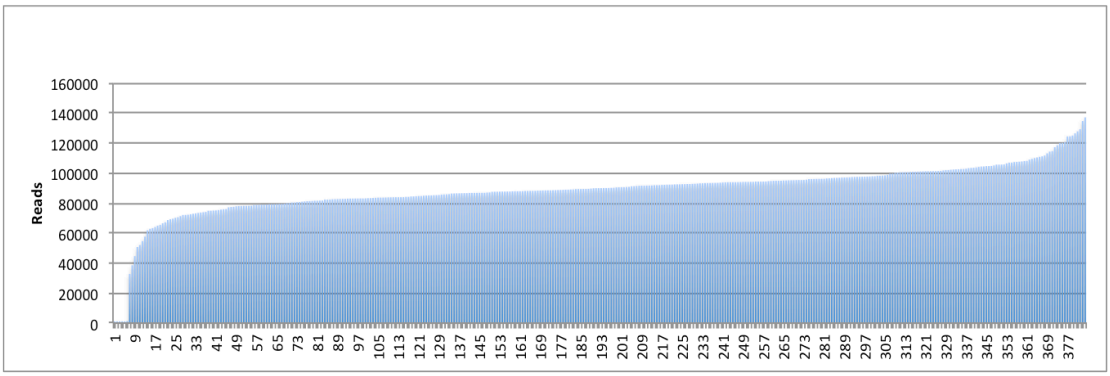
barcodes in this pilot study were attached via the initial amplicon PCR, and a second PCR step was performed to attach the bidirectional primers). The 140bp library was run on 6 lanes, each lane a different concentration of library: 1pM, 3pM, 6pM, 9pM, 12pM and 16pM. Data was aligned to hg19/build37 of an indexed human genome using Novoalign and output in SAM format. SNP calling was performed in SamTools v0.1.18. GATK v2.1-9 and SamTools v0.1.18 were used for depth of coverage analysis and variant calling.

5.3.3 Pilot study Results

From the Barts 140bp dataset, lane 5 (12pM concentration library) was optimal for the number of aligned mapped reads, coverage and evenness of reads. For this lane, 76.7% of reads were over 80bp (33,068,066 total reads). For the CRUK 100bp run, 95.8% of reads were over 80bp (34,162,646 total reads). Quality score distributions were generated using FastQC v0.10.0 software (www.bioinformatics.babraham.ac.uk/projects/fastqc). The Phred score for Barts lanes 1-6 was between 22 and 26, however there was a much wider distribution compared to the CRUK 100bp sequencing data (Appendix III, Figure 1). Overall, barcode evenness across 384 samples was excellent. Nine of the 384 barcodes had a <50% median number of reads, and this included 4 water samples (Figure 5.2). The median number of reads per barcode for 375 post quality control samples was 90,021. A higher total number of aligned trimmed reads was observed in the 100bp sequencing dataset (Figure 5.3).

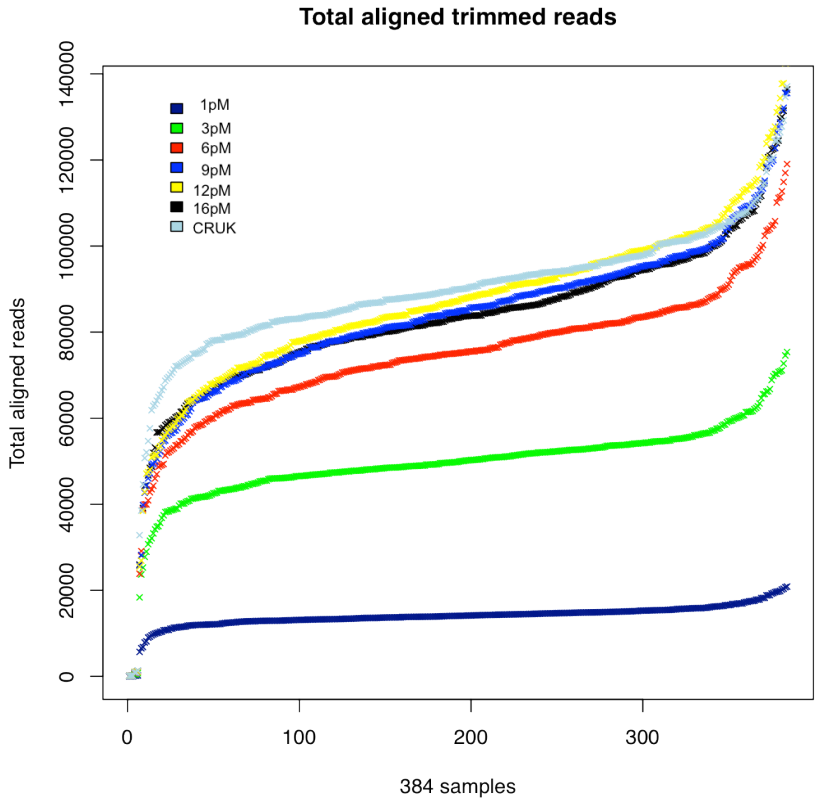
There were 44 SNPs with Immunochip genotypes and 339 of 367 samples had Immunochip genotype data. There were 14,851 calls made in both the Immunochip array and Fluidigm sequencing data, and 14,796 of these calls were concordant in both datasets, resulting in a concordance rare of 99.6%. This included one sample mix up (17 discordance calls) and 2 poor SNPs (10 and 27 discordant calls).

Figure 5.2: Total reads for 384 sample barcodes



100bp, 10bp index, sequencing data was used to produce this graph

Figure 5.3: Total aligned trimmed reads

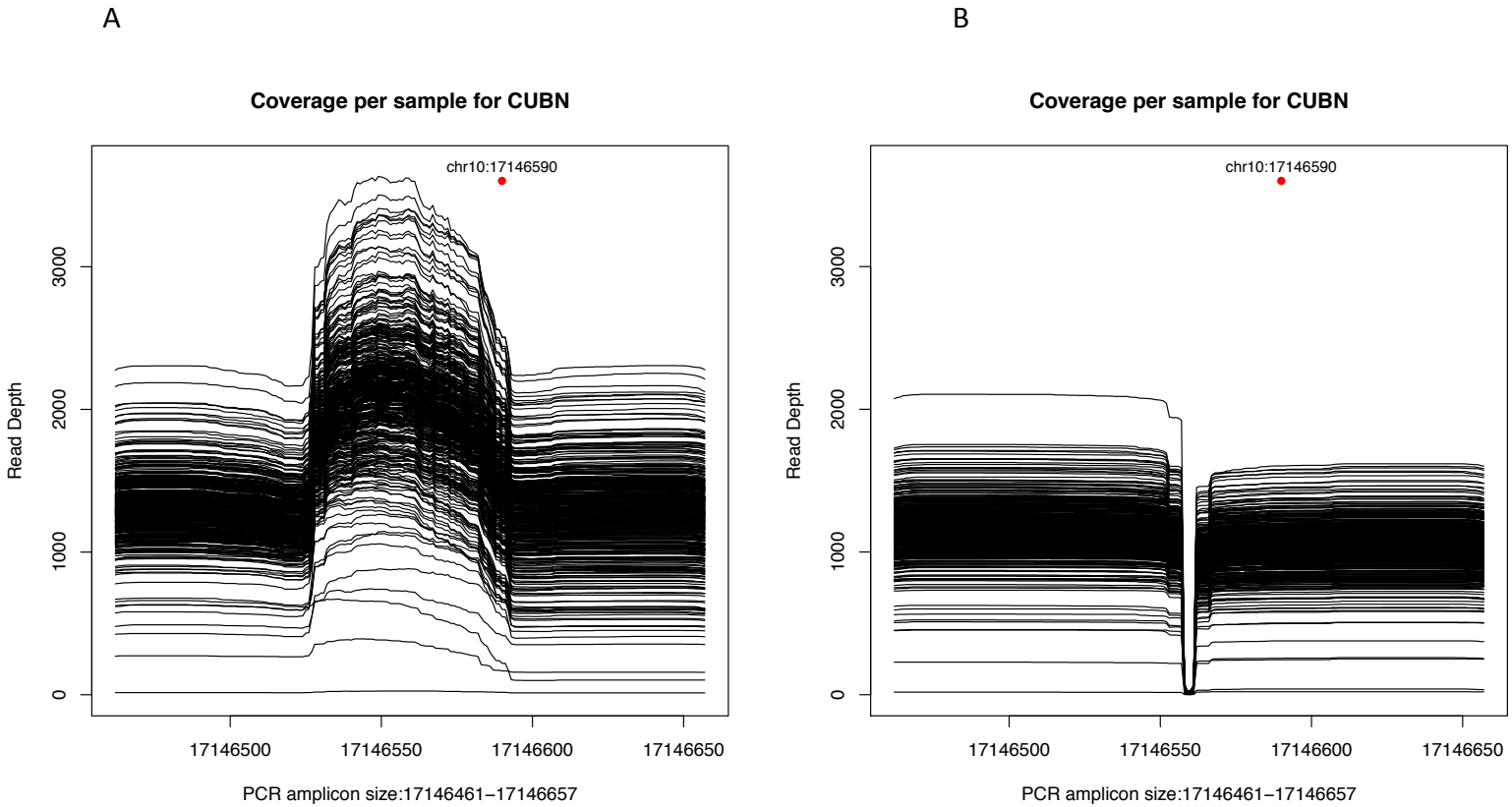


Concentratiosn ranging from 1pM to 16pM correspond to different library concentrations on one lane of the 140bp, 10bp index, sequencing run. CRUK corresponds to the 100bp, 10bp index, sequencing run.

5.3.3.1 Read depth analysis

It was important to establish depth and evenness of coverage per amplicon in order to assess if sequences at ends of reads were less reliable than the rest of the sequence. To computationally approach this question, a graph depicting amplicon coverage per sample was generated; lane 5 of the Barts data (140bp reads) was used to compare to the CRUK data (100bp reads). SamTools has a tool to generate pileup data from aligned bam files and the read bases at each position were extracted from each sample to produce graphs for each amplicon; figure 5.4 illustrates an example of typical coverage for one amplicon in *CUBN*. 140bp reads produced increased coverage where primers overlapped, but the 100bp read was slightly short to fully sequence the 196bp amplicon, however there was less error at the ends of reads.

Figure 5.4: Depth of coverage per sample for CUBN 196bp amplicon



140bp single end bidirectional reads (A) and 100bp single-end bidirectional reads (A) from 384 samples. Immunochip SNP highlighted in red.

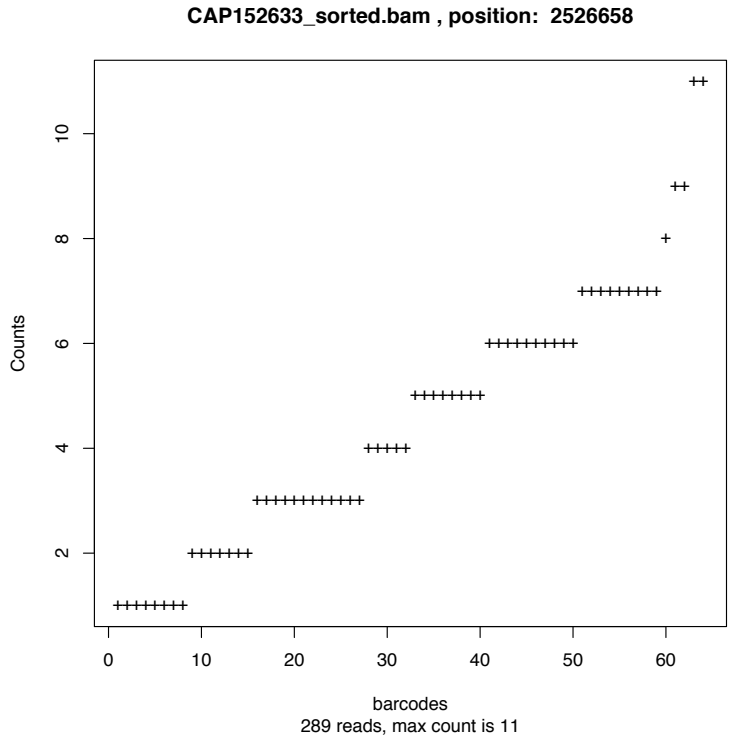
5.3.3.2 Clonality

One concern with PCR based library preparation methods is clonal or duplicate reads. These are multiple reads with the same orientation, start position and read length, and arise from PCR amplification Although PCR amplification increases the number of available molecules for sequencing, random errors can be introduced due to changes in the number and representation of template molecules. To recap, the exome sequencing data of pooled individuals had a high proportion of clonal reads e.g. 69.2% in pool 1, mainly attributed to a 2-step PCR (before and after hybridization, 18 and 20 cycles, respectively). This resulted in a large proportion of reads being discarded. It was therefore necessary to try and measure clonality in this pilot dataset, given the large scale multiplexing and 2-step PCR (one to amplify the region specific amplicons and a second to add the bidirectional primers).

A previously described method using degenerate bases as a molecular counter to estimate the number of template molecules in the amplified variant was applied here (Casbon, Osborne et al. 2011). The degenerate base, N, can incorporate an A, C, G or T. An amplified sequence displaying a mixture of bases at these 'N' reads indicates multiple variants have been sequenced and therefore amplified. The authors ligated 'N' bases onto fragments, however they can also be introduced via PCR, which was the method used here.

Three degenerate base tags (NNN) were added onto the target specific primers. A graph was produced by Vincent Plagnol to assess clonality. For each aligned bam file, reads were extracted from a given position and the counts of 'NNN' base tags was measured based on if they occurred more than three times the expected number (Figure 5.5). Out of 384 samples, only three samples exceeded the maximum "3-fold" threshold in the 100bp dataset, compared to 46 samples in the 140bp dataset. These results indicate that clonality is not a major downfall for this PCR methodology, especially with shorter reads.

Figure 5.5: Number of reads for position 2526658 for one sample



The number of counts is evenly distributed across the barcodes; the maximum degenerate base count is 11

5.3.4 Pilot Study Conclusions

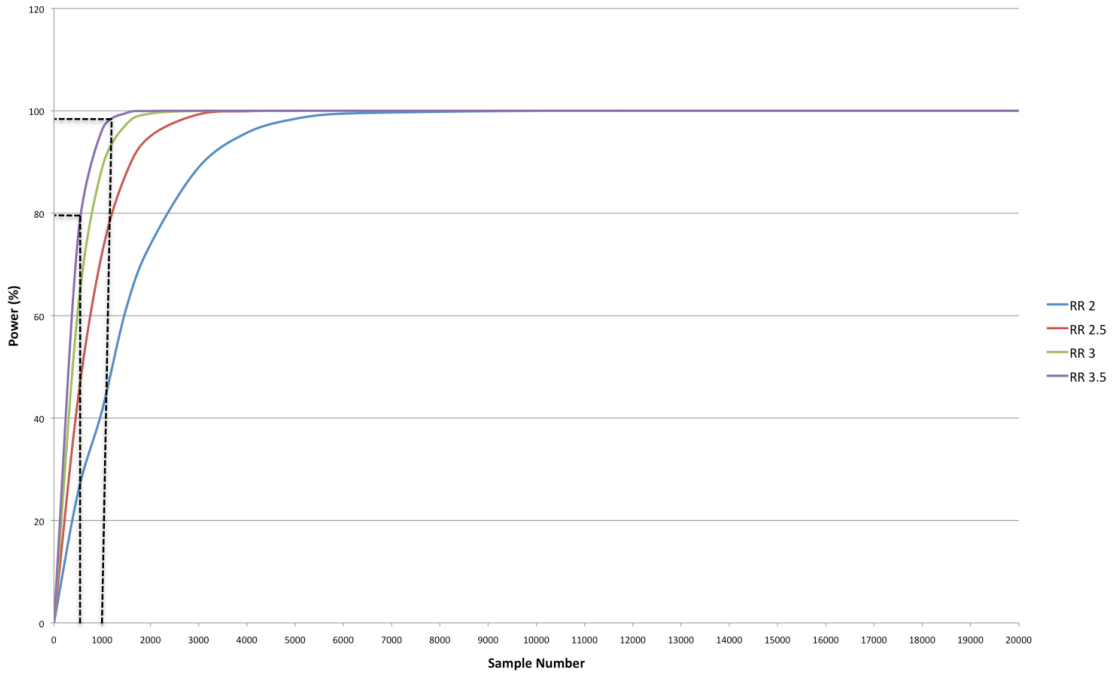
The outcomes from the pilot study proved helpful for assessing data quality using Fluidigm's Access Array IFC technology and high throughput sequencing on an Illumina platform. Overall, the data produced was at optimal coverage for variant calling and issues with clonality were bypassed with this technology, so the degenerate base tags were not added to the primers for the main project described later. It was evident that longer read lengths to sequence 150-200bp amplicons on the GAII*x* did not provide great read quality at the 3' end. Surprisingly, cluster density seemed to make no difference to the 3' base call quality. In addition, more reads over 80bp (95.8%) in the 100bp dataset were observed compared to reads over 100bp in the 140bp dataset (52.9% for lane 5). Furthermore, the barcodes were much more evenly distributed with 100bp

reads. The following conclusions can be made from this study: a) 100bp read length is optimal for multi-sample sequencing; b) degenerate base tags are not required as numbers of duplicate reads do not account for a large amount of the sequencing data; c) the array technology works well for large-scale amplicon PCR of >384 barcodes; d) Fluidigm's barcodes work well and at the time of completing this pilot they had designed an extra 1,152 barcodes for use, totaling 1,536 barcodes for multiplexing into one sequencing library, incorporated via a second PCR step. The bidirectional sequencing protocol was not yet optimized for use with the Illumina HiSeq 2000, which was the platform of choice for the main study. Instead, paired-end reads were performed, eliminating the PCR step required to attach bidirectional primers.

5.4 Power Considerations

A power calculation was performed to ascertain how many samples would be required to locate a variant with OR's (or relative risks) between 2 and 3.5 using an online genetic power calculator (Purcell, Cherny et al. 2003). Statistically, 1000 cases and controls would give almost 100% power to detect associations at a false positive rate of α=0.05 for an allele frequency of 0.5% and OR of 3.5 in a multiplicative model (Figure 5.6). Here, the risk allele frequency has been kept constant at 0.5%. Another calculation was based on an OR of a known rare variant in *NOD2*, G9068R (G>A). This marker has a genotype relative risk of 3.05 and a risk allele frequency of 0.12% in controls. The allelic OR's and population frequency of the minor allele, A, were calculated based on Hardy Weinberg, and equated to 1. The risk relative to the general population for individuals with genotypes GG, AG and AA for the rare marker was 8.86, 2.9 and 0.95, respectively. These numbers were used to calculate genetic power based on 2,304 disease samples and 2,304 controls. With a risk allele frequency of 0.5% and genotype relative risks based on G9068R, 2,304 cases and 2,304 controls gives 85% power to reach a *p* value of 0.001. Based on this, an overall 4,608 case-control sample size would be sufficient to find a high relative risk variant in CD, with a MAF of 0.5%.

Figure 5.6: Power calculation with increasing odds ratios and 0.5% risk allele frequency



The effect of sample size on power for variants conferring relative risks between 2 and 3.5. Assumptions: multiplicative model, disease prevalence 1%, high risk allele frequency 0.5%, case:control ratio 1:1. RR = Relative Risk. First dashed line indicates power for 500 cases and 500 controls. Second dashed line indicates power for 1000 cases and 1000 controls.

5.5 Experimental design and sample set

Fluidigm designed PCR primers for all RefSeq exons of 24 candidate genes totaling 506 amplicons containing exonic sequences (Table 5.1). Amplicons were selected to be 150-200bp in size. The design covered all exons, excluding any 5' or 3' un-translated regions. 21 out of the 24 target genes had 100% total coverage of all exon amplicons. There was minor primer design dropout at *MALT1*, *MAP4K2* and *IL12RB1,* however they still had a total coverage of 98.8%, 99.3% and 97.99%, respectively. The total length of (overlapping) amplicons was 96,581bp, 68,494bp with primers removed (and overlapping) and 55,807bp

unique (non overlapping and primers removed). One Fluidigm Access Array was intended to multiplex PCR 48 samples with 506 primer pairs (11-plex assay per well). The sample set consisted of 2,304 coeliac samples and 2,304 matching population controls, including negative (water) controls. Samples discordant with Immunochip genotypes and/or with known gender or genotype mismatch issues from previous GWAS were excluded (Trynka, Hunt et al. 2011). Samples with known duplicates or relatedness (as distant first cousins) were excluded.

Table 5.1: Candidate genes for targeted amplicon resequencing

Gene	Analysis selected from	cDNA size*	Exons*
ACOT8	Linkage	1,168bp	6
ARHGAP25	Linkage	2,979bp	11
C1QBP	Shared in familial exomes	1,169bp	6
CD180	Case Control/Shared	2,726bp	3
CD1C	Case Control	1,435bp	6
CERK	Case Control	4,450bp	13
CRLF3	Case Control	2,917bp	8
EBI3	Case Control	1,128bp	5
EPAS1	Linkage	5,160bp	16
GRM4	Linkage	3,879bp	10
HAS1	Shared in familial exomes	2,087bp	5
IFNW1	Case Control	1,514bp	1
IKZF3	Case Control	9,667bp	8
IL12RB1	Shared in familial exomes	2,100bp	17
KCNJ16	Linkage	4,002bp	5
MALT1	Linkage	8,789bp	17
MAP4K2	Shared in familial exomes	2,955bp	32
NLRC4	Linkage	3,581bp	10
RAF1	Shared in familial exomes	3,300bp	17
TNFRSF10A	Shared in familial exomes	2,714bp	10
TNFRSF13B	Shared in familial exomes	1,357bp	5
TNFRSF21	Segregation	3,595bp	6
TRAF4	Shared in familial exomes	2,921bp	7
TULP1	Linkage	2,162bp	15

*Information taken from Ensemble genome Browser, release 71

5.5.1 Laboratory method

Chapter 2, section 2.7 details the methods used for Fluidigm amplicon PCR. In brief, 50ng genomic DNA was PCR amplified in a multiplexed Fluidigm Access Array microfluidics system. One microfluidic array was used for multiplexed PCR of 48 samples (loaded on the left-hand side of the array). PCR primers for 506 PCR reactions were pooled up to 11-plex per well in 48 primer pools, which were loaded on the right-hand side of the array. Individual per sample per primer pool PCR reactions took place in 35nl reaction chambers, which took place in the microfludics chamber in the centre of the array. After a two-hour amplification and harvest of products, PCR amplicons from a sample were pooled and barcoded with one of 1,536 unique 10bp sequences (Fluidigm unidirectional sequencing protocol). An equal number of cases and controls were combined to create one 1,536 multiplex library (1μl per sample). Three libraries were generated in total. Libraries were initially sequenced on an Illumina MiSeq for quality control of individual barcodes and to optimise loading concentrations and cluster density targets for Illumina HiSeq sequencing. Libraries were then sequenced one per lane using 101bp paired-end reads and a 10bp index read on the Illumina HiSeq 2000 at NIHR GSTFT/KCL Biomedical Research Centre at Guy's Hospital. Individual samples were demultiplexed by Ilumina CASAVA software, allowing zero mismatches per 10bp barcode.

Sanger dideoxy sequencing was performed as in Chapter 3, section 3.5.2.1. All samples with rare variant allele genotypes, and a control sample, were sequenced for 27 sites selected.

5.5.2 In silico methods and quality control steps

Chapter 2, section 2.9.2 contains detailed bioinformatic steps performed on the 4,608-sample dataset. In brief, data analysis processes from 9,216 fastq files included: 1) PCR amplicon trimming using a modified version of Btrim software (Kong 2011), 2) read mapping with Novoalign to hg19/build37 of an indexed reference genome, 3) local realignment around known (1000G) and sample

level novel indels, 4) base quality score recalibration, 5) SNP and indel calling, and 6) variant annotation. Steps 3 to 6 were performed using GATK 2.4-7. The variants used were restricted to sites that passed standard GATK filters to eliminate SNPs with strand bias, low quality of read depth, homopolymer runs and SNPs near indels. Variants with an average depth >20 and a quality score >80 were required. SamTools v0.1.18 was also used to process data. SNP genotypes were called at all 68,494 bases of amplicon sequence. Non-reference genotype sites were identified across all samples and VCF files containing polymorphic variant sites and samples were combined and loaded into a project for use with PLINK/SEQ v0.09 software. Annotation in PLINK/SEQ was performed with GENCODE V14 gene definitions (Howald, Tanzer et al. 2012). Coding variants were identified as present in coding regions, and rare functional variants were identified based on nonsense, splice, esplice (splice site in the first or last two intronic bases), frameshift indel, codon indel (3n indel), readthrough, and start lost predictions. PLINK/SEQ v0.09 was used to perform all single variant and gene-based association analyses and for determination of TiTv statistics.

Quality control steps of the combined SNP and indel VCF file in the PLINK/SEQ project included the removal of: water samples (negative controls), samples with low call rates across all SNVs, SNVs with low call rates across all samples. The initial PLINK/SEQ project contained 2,292 polymorphic variants and 4,608 samples. A SNP and indel call rate of 97.7% and individual genotyping call rate of 97% (across all SNPs and indels) was applied (thresholds determined by inspection of call rate plots in Appendix III, figures 3 and 4). All heterozygous calls were required to have an allele balance between 25% and 75%, calculated by the division of alternate allele depth over total allele depth for a site (Appendix III, Figure 5). The mean allele balance at all heterozygous sites was 0.49±0.12. The mean ± two standard deviations was 24% and 73%, similar to the 25%-75% allele balance used here and in Hunt et al (Hunt, Mistry et al. 2013) (Appendix IV), and the same number of variants was removed when both filters were applied (52). Another recent study used variants with an allelic balance between 30% and 70% (Lim, Raychaudhuri et al. 2013); this would have

removed 91 variants from the dataset, however the called variants had a genotyping rate of >99% with 25% -75% filters so these limits were kept. GC content can explain why low depth was observed at some, but not all, sites. For example, a cluster of sites with low heterozygote allele depths in one amplicon of *EPAS1* and three amplicons of *HAS1* contained 69%, 73%, 67% and 70% GC content, respectively (Appendix III, Figure 5).

5.6 Results

On the basis of MiSeq 50bp single-end sequencing results for library quality control, three libraries contained excellent barcode coverage across 1,536 10bp sequences, with 99.6% of the 1,536 barcodes producing pass-filter read numbers. These were between 0.013% and 0.13% of total pass filter reads per lane. Most failing barcodes were water (negative control) samples. For three HiSeq 101bp, 10bp index, paired-end sequenced libraries, >93% reads passed filter with on target cluster densities between 640 -775 k/mm2 (Appendix III Table 1 and Figure 2). Amplicon evenness was good with many genotypes requiring down-sampling of 250 bases per site per sample (Appendix III, Figure 6). A filter of >20 mean depth per sample was applied to call a variant. Amplicons were visually inspected on UCSC tracks, and those with <20 mean coverage were removed. 3.47% of 55,807 unique bases had <20 mean depth per sample and were all accounted for by 18 amplicons that failed PCR. 13 out of the 18 failed amplicons had high GC content (between 63 and 89%).

The high coverage data enabled stringent filtering on call rate per sample, per variant site and on allelic balance, as described in section 5.5.3. After quality control and removal of excess related individuals, ethnic outliers and duplicates, the final dataset comprised 4,478 phenotyped individuals (disease cases and controls). 1,335 unique variants with a genotype call rate of 99.98% were discovered. The genotyping call rate included reference homozygote and non-reference genotypes. Of these, 1,200 variants were rare (MAF in 2,230 post quality control controls, <0.5%). 502 variants were observed in published datasets (dbSNP137 containing all 1000G pilot data plus phase 1 low coverage

sites and National Heart, Lung and Blood Institute (NHLBI) exome data from 6,503 samples) and 833 variant sites were novel. 99.98% of all sample genotype calls had a read depth >40 and 97.4% had a read depth >100.

The number of coding variants per gene was assessed and limited to heterozygous variants only. Here, a coding variant is defined as one that is present in the coding region; therefore, silent variants have been included in the count (Table 5.2). Of the 1,335 variants, 939 were in protein-coding regions of 24 genes and of these 91.7% were rare (MAF in 2,230 controls, <0.5%). 60% of all coding variants were novel when compared with published datasets (as above). No common or low frequency variants were seen at novel sites (mean MAF 0.00139%). Overall, 60 rare LoF variants (nonsense, codon indel, frameshift, and splice site; based on GENCODE v14 annotation) were identified across 20 genes; four genes harboured no such variants.

Data quality was confirmed by a number of steps. One control sample was genotyped 42 times (on different 48-sample microfluidic chips); the genotype call error-rate was two non-consensus genotype calls of 1,335 called genotypes (0.0018%). A quality control step measuring TiTv ratios for expected human mutation types was 2.99 (3.18 singletons) for coding-region variants, 2.86 (3.13 for singletons) for rare variants, and 2.69 (2.90 for singletons) for novel variants. For novel coding-region variants the TiTv ratio was 2.78 (2.89 for singletons). Sanger sequencing validation analysis was performed on all nonsense (17) and frameshift (11) variants. There was one variant that failed PCR and one frameshift indel and one nonsense variant that were false positive (false-positive rate = 7.4%).

Table 5.2: Number of coding, rare and loss of function variants across 24 candidate genes

Gene	Number of variants in coding regions	Number of rare (MAF<0.5) in coding regions*	Number of rare (<0.5) and LoF**
ACOT8	29	27	5
ARHGAP25	40	34	2
C1QBP	9	8	2
CD180	48	43	4
CD1C	28	26	3
CERK	54	48	4
CRLF3	23	20	2
EBI3	28	25	2
EPAS1	59	55	3
GRM4	69	65	0
HAS1	61	56	1
IFNW1	17	16	0
IKZF3	29	27	1
IL12RB1	60	52	4
KCNJ16	36	35	3
MALT1	21	20	3
MAP4K2	34	33	2
NLRC4	64	61	3
RAF1	27	26	2
TNFRSF10A	46	42	5
TNFRSF13B	42	34	4
TNFRSF21	39	38	0
TRAF4	30	28	0
TULP1	46	42	5

* MAF as defined in controls. **Loss of function excludes synonymous and silent variants.

5.6.1 Association and gene-burden tests

A first attempt to identify any low frequency or rare variants of larger effect was performed for each coding-region variant in a Fisher exact single-variant association analysis. 135 variants common in controls were removed from the test (>0.5% MAF). A significant *P* value of 6×10^{-5} was chosen to account for multiple testing on 939 rare coding variants. No single SNP associations were

observed (the highest *P* value was 0.012). Gene based tests were subsequently performed in PLINK/SEQ on all coding variants across 24 genes. A gene based C-alpha test allowed for both risk and protective effects for rare functional variants. A sequence kernal association test (SKAT; a variance-component test that aggregates individual score statistics by assigning weights for each SNP to perform (Wu, Lee et al. 2011)) and tests to identify excess rare variants seen in cases, collectively (Burden test) and uniquely (Uniq test), were also performed. Rare functional variants included in the tests were defined as <0.5% in 2,230 controls and predicted nonsense, frameshift, codon indel and splice site annotation. Here, a Bonferroni *P* value of $<1 \times 10^{-3}$ was selected based on the number of transcripts tested, and not the number of genes, as some genes had multiple transcripts. No significant *P* values were observed in any test for novel or known variants. Table 5.3 shows genes with top five *P* values across all gene-based tests.

Table 5.3: Top five *P* values for multiple rare variant gene-based tests across all protein-coding variants (novel and known) in 24 candidate genes

Gene	Transcript	Rare variant test	Number of variants in test	Test statistic p value
CERK	NM_022766	C-Alpha	48	0.022
ARHGAP25	NM_001007231	C-Alpha	34	0.118
HAS1	NM_001523	C-Alpha	56	0.119
IL12RB1	NM_005535	C-Alpha	52	0.229
TNFRSF13B	NM_012452	C-Alpha	34	0.275
CERK	NM_022766	SKAT	48	0.002
ARHGAP25	NM_001007231	SKAT	34	0.096
HAS1	NM_001523	SKAT	56	0.126
IL12RB1	NM_005535	SKAT	52	0.188
CD1C	NM_001765	SKAT	27	0.263
EPAS1	NM_001430	UNIQ	55	0.004
CD1C	NM_001765	UNIQ	27	0.044
HAS1	NM_001523	UNIQ	56	0.092
IFNW1	NM_002177	UNIQ	16	0.140
RAF1	NM_002880	UNIQ	26	0.229
EPAS1	NM_001430	Burden	55	0.007
ARHGAP25	NM_001007231	Burden	34	0.167
TNFRSF21	NM_014452	Burden	38	0.234
CD1C	NM_001765	Burden	27	0.240
TNFRSF10A	NM_003844	Burden	42	0.262

5.7 Chapter Discussion

The follow up study in this chapter has attempted to identify rare protein-coding mutations in 24 candidate genes selected from exome sequencing data of 75 coeliac individuals from 55 multiple affected families. Candidate genes were chosen based on if they harboured any rare variants shared in familial exomess (with the assumption that closely related affected individuals share rare risk variants), if they were in linkage peaks or if more variants were observed in cases compared to controls in any particular gene (gene burden test). The study was intended to prove or disprove the working hypothesis that rare mutations of large effect size in CD may account for the missing heritability of disease, and these variants can be found through NGS of candidate genes.

The investigation of the role of rare variants in complex traits (Gibson 2011) has led to many studies determining their distribution across the genome, their phenotypic affects and how to apply statistical aggregate tests to quantify their presence in a disease population (Asimit and Zeggini 2010; Liu and Leal 2012). The relevance of rare variants has been widely demonstrated, for example, in a quantitative trait where many rare nonsynonymous SNPs unique to the obese population have been discovered (Ahituv, Kavaslar et al. 2007), and in complex phenotypes such as autism and Parkinson's disease, where rare structural variants play a significant role (Stankiewicz and Lupski 2010). One of the main purposes of elucidating the role of rare variants in functional regions of the genome is to pinpoint a specific gene(s) in which the presence of variants with a low MAF in the general population is attributable to disease risk in a disease cohort. One key concern is having the statistical power to detect these variants, which have an overall low population frequency but are in abundance in the human population, due to recent population expansion (Coventry, Bull-Otterson et al. 2010). The 1000 Genomes project found that for rare variant testing to pinpoint disease associations, there is a significant reduction of power due to a large number of variable sites and a lack of sharing of these variants amongst diverge populations. So in essence, variants with any functional impact are rare and have a higher population divergence (Gravel, Henn et al. 2011). However,

with enough statistical power, testing genes rather than multiple alleles as units for an association test can identify meaningful associations (Kryukov, Shpunt et al. 2009). The experiment here was successful in that many rare mutations were found (91.7% of all coding SNVs were rare) and the proportion of novel and rare variants were similar to those found in other published datasets (Nelson, Wegmann et al. 2012; Tennessen, Bigham et al. 2012). The data also reflects the operation of purifying selection as all nonsynonymous substitutions were skewed toward a low MAF (0.09%), similar to other findings (Cargill, Altshuler et al. 1999; Kryukov, Pennacchio et al. 2007; Zhu, Ge et al. 2011).

The Fluidigm multiplex PCR method used in this study was excellent in providing the coverage required to confidently call a variant. Overall, the 1,536 barcode pooling provided even coverage across all samples and amplicons, with only 3.5% of amplicons failing PCR, mainly due to GC rich content. The 2-step PCR (target specific amplification followed by barcode PCR) did not introduce clonality into the dataset, possibly due to the CoT PCR method employed by Fluidigm to normalize PCR reactions (Mathieu-Daude, Welsh et al. 1996). By applying CoT PCR with such a large number of amplicons in small PCR reactions, under-performing amplicons have a method to catch up with those that are better performing. Compared to error-prone low coverage sequencing where genotype calls cannot be confidently made (Navon, Sul et al. 2013), highly multiplexed PCR amplicon sequencing proved to be an efficient method to gain the depth needed at multiple variant sites in a large case control cohort. Miscalling errors can be prevalent in sequencing data at individual genotypes, especially when attempting to discriminate rare variants, which can easily be perturbed due to their low MAFs. In addition, particular NGS errors are systematic according to the platform used, for instance A>T miscalls are most common in Illumina data (Bravo and Irizarry 2010), and the quality of reads are significantly lower in later cycles, possibly caused by incomplete extension of the template (Metzker 2010; Nakamura, Oshima et al. 2011). The Illumina GAII*x* 100bp data from the pilot study resulted in less error at the ends of reads compared to 140bp data, so 101bp paired end (10bp index) sequencing runs were chosen for the main study. The final dataset here contained only minor

miscalls (7.4% false-positive rate) and all Sanger validated SNVs and indels had the same alleles in Sanger-sequencing assays as in the high throughput sequencing data. One SNV annotated as a frameshift indel, was in fact validated as a triallelic SNP. Firstly, the fact the SNP was reported as a triallelic variant and not a bi-allelic SNP shows improvement in the calling methods now available (i.e. GATK). This is also true for the other triallelic SNVs (34 in total) observed in the data. It is not uncommon to see three nucleotides at one site, and mutational mechanisms have been proposed that could potentially generate an excess of triallelic sites. One explanation could be due to an elevated mutation rate at a CpG site of at least two pathways (e.g. C-T or C-A) and CpG sites have shown to be elevated in exons compared to overall occurrence in the whole genome (Saxonov, Berg et al. 2006). However, one study proposed that instead of triallelic SNPs occurring at particular sites, they can incur during recombination within the same individual (Hodgkinson and Eyre-Walker 2010). Overall, the consequence of not detecting the unknown allele of a triallelic SNP can be serious when the unknown allele carries a disease risk (Huebner, Petermann et al. 2007). Apart from Sanger validation, other data quality steps were necessary; assessment of pass filter reads from water negative controls and calculation of error rate of one positive control sequenced multiple times, but PCR amplified on different multiplex arrays. Combined data quality results highlighted the high specificity and sensitivity of the dataset enabling confident genotype calls at variant sites across the sample set.

It was noted that with 4,478 (post quality control) sample set, single-variant association mapping of all coding rare variants was low powered for the number of variants in the test and their associated low MAFs resulted in the test being numerically unstable to analyze each variant independently (Bansal, Libiger et al. 2010; Morris and Zeggini 2010). Therefore gene-based tests, in which multiple rare variants in the gene region are jointly analyzed to aggregate all signals, were performed to better detect the combined effects of multiple variants, given the evidence that multiple rare variants can have a collective effect on disease risk (Cohen, Kiss et al. 2004; Fearnhead, Wilding et al. 2004). These tests also reduced the effects of multiple testing, as the test was based

on each gene transcript rather than each individual variant. Given that no significant associations were identified, one fair conclusion is that the 4,478 sample dataset lacked statistical power to achieve a significant rare variant association at a candidate gene. For the gene with the highest *P* values (*P*=0.004 in a uniq case-control allele test and *P*=0.007 in a burden test), *EPAS1*, a larger sample size may possibly have pushed the *P* value to a significant association at $P<1 \times 10^{-3}$, indicating that there is a significant excess of alleles in cases compared to controls. Even when correcting for the number of genes, rather than the number of transcripts tested (as some transcripts for the same gene had the same number of variants), no significant associations were present at $P<1 \times 10^{-3}$.

A major difficulty with complex disease is selecting genomic locations to focus on, compared to, for example, an X-linked disease where all X-chromosome genes can be easily resequenced (Tarpey, Smith et al. 2009). In this study, it was necessary to balance the cost of resequencing with the number of genes that might carry a rare risk variant. The candidate list was developed through multiple strategies using exome capture data, where not all the exome will have been captured. Given that only a small fraction of the genome was interrogated (amplicons only cover small parts of the chromosomes), all samples were directly sequenced so the sample set was utilized in the best way possible to search for rare variants in investigative regions. The intention was that sequencing could lead to identification of enough rare variants, in one or more genes, that could explain a substantial proportion of the missing heritability. However, even if a single gene with a large effect size in CD had been found through the chosen analytical strategies, the therapeutic advantage would have carried a larger pay off. For example, *BRCA1/2* mutations do not necessarily explain much heritability of breast cancer, but there is diagnostic value for treatment (Couch, Wang et al. 2013).

On the basis of the results, there is lack of support to prove the rare variant-common disease hypothesis in subjects where there is familial clustering of disease. These findings support a recent resequencing study of 25 GWAS risk genes from six autoimmune diseases in 42,000 subjects that concluded rare

coding mutations play a negligible role in the autoimmune diseases under investigation (Hunt, Mistry et al. 2013). The role of rare variants in common disease is further discussed in Chapter 6: Discussion.

5.8 Chapter Conclusions

The points below conclude the findings from the research in this chapter:

1. Fluidigm Access Array multiplex PCR technology was excellent in providing the necessary coverage required for highly multiplexed amplicon sequencing.
2. The high quality dataset was acquired by applying high individual call rates and SNV genotyping call rates, along with other data quality filters.
3. Many rare variants in 4,478 post quality control samples were observed in protein coding regions of 24 candidate genes, of which 60 SNVs were rare (MAF <0.5%) *and* LoF, according to GENCODE v14 annotation.
4. Various gene-based methods to test the effects of multiple rare variants per gene were performed, but no significant associations were observed in any candidate gene.
5. For *EPAS1* (*P*=0.004 in a uniq case-control allele test and *P*=0.007 in a burden test) testing the variants a larger sample size is required to confirm any significance with disease or not.

Chapter 6

Research Discussion

6.1 Research background and summary of findings

The spectrum of genetic variation in human diseases ranges from common variants of combined weak effects to rare variants with modest to strong effects on the disease phenotype. Rare, high-risk, monogenic variants have classically been determined through linkage and positional cloning analysis in Mendelian pattern diseases (Kerem, Rommens et al. 1989; Riordan, Rommens et al. 1989), and more recently exome sequencing, whereas a large fraction of common variants have been investigated by GWAS in common complex diseases (Hirschhorn, Lohmueller et al. 2002), made possible by the completion of the HapMap project (Consortium 2003; Thorisson, Smith et al. 2005). The CDCV hypothesis was rooted in descriptions of the allelic spectrum of disease, outlining the totality of variations contributing to disease: low penetrance (disease is expressed in a minority of individuals with the phenotype), high penetrance, common and rare variations (Reich and Lander 2001). Since variants of relatively low penetrance but with common population allele frequencies (that are shared between multiple individuals) have not explained much of the heritability of disease risk (due to genetic effects), focus has been shifted toward locating multiple rare variants with high penetrance that could explain the majority of genetic susceptibility to common disease (the CDRV hypothesis) (Schork, Murray et al. 2009). Studies have firmly proposed that amino-acid altering (nonsynonymous) mutations generally have a rare population frequency compared to silent substitutions (Cargill, Altshuler et al. 1999), due to negative selection acting upon deleterious mutations preventing them from reaching high allele frequencies, producing an excess of rare variation (Fay, Wyckoff et al. 2001; Bustamante, Fledel-Alon et al. 2005). This excess has also been attributed to recent population expansion (Williamson, Hernandez et al. 2005) with one study predicting 30-42% of nonsynonymous (amino acid altering) mutations are moderately deleterious after accounting for demographic effects in African and European populations (Boyko, Williamson et al. 2008).

A major advance in the field of genetics in the last five years has been the creation of technology that allowed capture and sequencing of the entire protein-coding region of the genome by companies such as NimbleGen and Agilent. Since the technology for exome sequencing became available, researchers were able to carry out hypothesis-free studies using sequenced variants in the exome, similar to GWAS studies using genotyped markers. Filtering approaches and investigating novel variants in genes common to disease individuals led to many studies publishing findings, some of which enabled a genetic diagnosis in rare Mendelian traits (Hoischen, van Bon et al. 2010; Krawitz, Schweiger et al. 2010; Pierce, Walsh et al. 2010; Wang, Yang et al. 2010). Exome studies in mental disorders, such as autism, schizophrenia and mental retardation, have also located rare CNVs in gene risk pathways (Stefansson, Rujescu et al. 2008; Helbig, Mefford et al. 2009; Pinto, Pagnamenta et al. 2010), and in combined consanguinity and homozygosity mapping studies (Caliskan, Chong et al. 2011). The success of these studies are also owed to the expansion of sequencing technologies for deeper resolution of novel and known genetic variants (Shendure and Ji 2008; Metzker 2010; Nekrutenko and Taylor 2012), where the 1000G project has been at the forefront of uncovering variation of at least 1% frequency in major population groups (Abecasis, Altshuler et al. 2010; Abecasis, Auton et al. 2012).

The primary aim of the research in this thesis was to locate rare variants predisposing to CD risk, through target capture of protein coding regions and high throughput sequencing, with evidence of positive findings from candidate gene studies in other genetic diseases (Yeo, Farooqi et al. 1998; Kotowski, Pertsemlidis et al. 2006; Raychaudhuri, Iartchouk et al. 2011). To recap, whole-exome sequencing (Chapter 3) was selected as a method to identify any rare (<5% MAF) mutations that might directly affect gene/protein function in CD individuals selected from 55 multiply affected (>2 affected per family) disease families. A family-based sample set was used to detect a Mendelian pattern of inheritance in an otherwise complex genetically heterogenous disease (Cardenas-Roldan, Rojas-Villarraga et al. 2013). Because of the highly heritable nature of CD (Nistico, Fagnani et al. 2006), if there were a high risk variant

causing severe phenotypes it should be present in families. Any highly penetrant rare mutations found might explain the ~60% of missing heritability in CD, if associated with disease in a large case control sample set.

Early on, it was recognized that the main caveat to the experiment was many thousands of candidate variants (average 15,601 protein coding mutations per sample) were identified, creating a challenge in locating true disease causing mutations. The intention was to apply various strategies to the dataset generated from a newly released capture method (NimbleGen in-solution exome capture, released in 2010) requiring new strategic applications, from segregation analysis to gene-based tests. No conclusive variants were found to be segregating with disease in >2 generation families with >2 affected individuals. Gene-based analysis identified *CUBN* as a potential candidate with three different mutations in three unrelated individuals, indicating rare mutations in this gene may carry a disease risk. This gene might be taken forward as a future experiment since it was too large to resequence here in a Fluidigm multiplex assay. Other candidate genes that carried more or less variants then expected in cases compared to controls (identified by a Fisher's exact test) were selected for resequencing. Another strategy to locate shared chromosomal regions with potential rare variants in coeliac families was NPL linkage analysis. Significant linkage (LOD >3) was not observed in 12 pedigrees, however many peaks with LODs of >1 were present. Since these peaks highlighted shared chromosomal regions, exome variants under the peaks were assessed and Sanger sequenced in the entire family to confirm the variant's presence on a shared haplotype. Eight genes were taken forward for resequencing. Interestingly, a finding that is also conclusive across all coeliac studies, was the presence of common HLA risk allele types segregating in 12 linkage pedigrees, some of which came from persons married into the family and therefore not ancestral.

To study further rare variation in candidate genes (n = 24), gene-based tests were performed on the candidate gene sequenced dataset in 2,304 cases and 2,304 controls. Many missense and silent mutations were observed in exons of genes, and only a handful of LoF mutations. 91.7% of coding variants were rare,

similar to findings in a recent rare variant autoimmune disease study (Hunt, Mistry et al. 2013). No mutations were significantly associated with disease risk or protection, however a larger case control sample set is required to test for any significant association in *EPAS1* (*P*=0.004 in a uniq case-control allele test and *P*=0.007 in a burden test). *EPAS1* (also known as *HIF2A*) is a transcription factor expressed in endothelial cells and produces the hypoxia-inducible factor 2-alpha protein, which plays a role in the body's adaption of changing oxygen levels. A gain of function mutation in *EPAS1* has been implicated in familial erythrocytosis (Percy, Furlow et al. 2008). In a study testing the effects of hypoxia-inducible factors 1 and 2 (*HIF1* and *HIF2*) in diabetic mice, the authors found that glucose activates *HIF1* and *HIF2* in rat beta cells and both isoforms are activated in islets suggesting hyperglycaemia could induce pancreatic beta-cell hypoxia (Bensellam, Duvillie et al. 2012). Since many risk loci are common to both T1D and CD, *EPAS1* may also play a role in CD pathogenesis. Further studies of the possible function of this gene in coeliac patients must first confirm any significant association followed by a sequencing replication study of the entire exonic region in more samples. Even a meta-analysis combining known T1D and CD GWAS and fine mapping genotypes with imputation of rare variants from sequenced variants in this study could highlight an association in *EPAS1* using a large case control sample size, as observed in other autoimmune diseases (Zhang, Yan et al. 2013; Zheng, Yin et al. 2013).

A recent study applied imputation of rare whole genome sequenced variants into >95,000 Icelandic genotyped individuals and found that a rare nonsense mutation in *LGR4* associated with low bone mineral density in osteoporosis also had an increased risk in squamous cell carcinoma and biliary tract cancer (Styrkarsdottir, Thorleifsson et al. 2013). This study does highlight some key points: imputation from rare variants into tagged SNPs is not completely accurate, but the authors did apply familial imputation by way of increasing accuracy as one would expect related individuals to have the same genotypes (they used familial sequenced variants to impute into un-genotyped relatives); the study covered all genes genome-wide compared to 24 candidate genes used here and yet only discovered one rare (0.1%) mutation, and loss of

heterozygosity was only assessed in four biliary tract cancer variant carriers, so more functional work is required to completely assess phenotypic effect in carriers. Since high effect rare variants are known in these cancer types, one would expect rare variant detection in such a big sample size, which may not be the case for a common disease, as observed for rare variants in *SIAE* where there was a lack of association across multiple common autoimmune diseases (Hunt, Smyth et al. 2012).

6.2 Effects of sample and experimental design

Choosing the correct sample set for exome sequencing is a key component in the experimental design. It is unfeasible for researchers to sequence the tens of thousands of individuals required for rare variant detection because they are present at such low frequencies. To maintain adequate costs and balance the potential of finding a result, the 75-sample set here was a combination of first and second-degree relatives from coeliac families. This is akin to selecting extreme phenotype samples for a quantitative trait (Li, Lewinger et al. 2011), and since pedigrees had a history of disease in at least two generations, the potential of finding a variant segregating in a Mendelian fashion was felt to be high. In addition, a previous study showed that relatives of coeliacs have a high risk of carrying silent CD, so using family samples can enrich for any disease causing mutations (Petaros, Martelossi et al. 2002).

Given these assumptions, in a family design for rare variant detection, the question that should be addressed is under what circumstance are rare variants with a disease risk expected to segregate in familial disease cases. A recent study found that as λ_s (defined as the ratio of disease manifestation if one sibling is affected, compared with prevalence of disease in the population) increased, the probability of co-segregation of the rare variant declined (Helbig, Hodge et al. 2013). The authors used the presence of the 15q13.3 microdeletion in probands with idiopathic generalized epilepsy as an example and noted that for a variant with an OR of 5, the probability that an affected relative carrying the variant was 58% when the λ_s was 50, compared to 82% when the λ_s was 2.

However, for moderate ORs (between 1 - 1.3), λ_s had little effect on the probability. Another study similarly highlighted that for a complex disease with a high λ_s sequencing unrelated individuals was preferable, but for diseases with small λ_s sequencing an affected individual with an affected close relative can be a powerful strategy for the identification of rare variants, based on a model using 2,000 affected individuals (Ionita-Laza and Ottman 2011). The λ_s of 10 for CD might not have an effect on the probability of an affected relative carrying a rare risk variant, according to the probability theory employed in Helbig et al. The strategy of sequencing one or two affected relatives per family (with an affected close relative) in this study was moderately powerful, according to the model outlined in Ionita-Laza et al, however a sample size of >75 was probably required to detect those shared variants, albeit at a cost of an increased number of variants to filter through. On the other hand, if the funding was available, it might have been advantageous to sequence every affected member of the family to increase power. Exome sequencing data in consanguineous eczema families has located multiple variants in a single known disease gene, *FLG*, observed in different affected individuals. No single variant was common to all affected cases revealing that rare variants are unique to the individual (David Kelsell, personal communication). If one gene was causal to CD risk, all affected individuals would have to be sequenced in order to locate all risk variants in the gene clustering in the family.

The control sample set is just as important as the case sample set. Recent experiments testing the outcome of a combination of rare variant tests using three different control datasets with three family structures (trios, enriched trios, where only one sibling is tested, and ASP) showed that using unrelated controls gave better power than using controls from family data (Preston and Dudbridge 2013 *in press*). No related control samples were exome sequenced in this study, but 200 unrelated exome controls from neurodegenerative phenotypes from the UK were used. Small differences in ancestry between cases and controls can still incur false positive results since rare variants have a restricted geographic distribution (Mathieson and McVean 2012), but given that similar false positive rates were observed in the initial discovery exome dataset

as in the candidate gene resequencing dataset (8.1% and 7.4%, respectively), which had a matching population control dataset of 2,304 individuals, the unrelated independent control set for the exome sequencing experiment was a good choice.

Linkage analysis in this thesis was performed to locate shared chromosomal regions in coeliac families in which segregating rare variants might be found. Several factors affect the power of a linkage test to produce a result: polymorphism of the marker, mode of inheritance of the disease and recombination distance between the disease locus and marker, all of which may have been factors in the analysis here. Using a multiplicative linkage model, it has been shown that diseases with a large λ_s often indicate genotype specific effects that are not well detected by linkage methods and multiple loci having a combined effect must be taken into account (Rybicki and Elston 2000). The authors also highlighted that IBD distribution amongst distant relationships (i.e. uncle-nephew) is dependent on marker allele frequencies, so although it is beneficial to use these relationship types, for a rare marker allele there could be a lack of IBD. The linkage study here used markers with MAF <0.2, so IBD sharing may not have been significant amongst the more distantly related individuals in the pedigree. There is also the issue of multiple phenocopies, defined here as the disease being acquired by different means, i.e. different underlying markers in affected individuals compared to other cases in a pedigree. In one study, phenocopy impact has been shown to have a significant effect when epistatic effects were included in a simulated disease model (Lescai and Franceschi 2010). Correcting for these interaction effects may significantly improve the linkage LOD score (Sung and Wijsman 2007).

6.3 Progression in exome studies

This work in this thesis was one of the earliest projects aiming to seek rare variants in the exome, and since then many changes have occurred surrounding exome sequencing. At the time of project commencement (2009), the cost of sequencing one exome was approximately £1,200, using NimbleGen exome

capture and Illumina sequencing. Companies such as 23andMe are now offering prices of $999 per exome, so the cost is not significantly different when taking library preparation costs into account. However, as the cost of NGS decreases, it is cost-effective to sequence multiple exomes, especially in a collaborative effort. The effectiveness of exome sequencing has initiated large collaborations employing the method for the discovery of novel genes contributing to disease phenotypes. The NHLBI Exome Sequencing Project is a combination of heart, lung and blood disorders from large, well-characterized populations in the United States with the aim of finding novel disease mechanisms. A plethora of studies have already been published (Regalado, Guo et al. 2011; Boileau, Guo et al. 2012; Emond, Louie et al. 2012; Krumm, Sudmant et al. 2012; Norton, Robertson et al. 2012; Tennessen, Bigham et al. 2012). Some studies evaluated all the rare coding variation in relation to functional impact, with similar findings to other coding variation studies (Norton, Robertson et al. 2012; Tennessen, Bigham et al. 2012), and one study used the data to detect CNV variation showing high correlation with whole-genome data when using exome-based genotyping to detect copy number polymorphisms (Krumm, Sudmant et al. 2012).

Similar efforts are established in the UK to detect new discoveries in a large European cohort. The UK10K initiative aims to sequence 6,000 exomes with an average 72x sequencing depth and 4,000 whole-genomes with an average 6x depth. Where the 1000G sequenced individuals from distinct geographic populations, the UK10K aims to provide a deeper resolution in a European cohort. The 6x low pass sequencing coverage has a resolution to detect more variants than the 4x depth used by 1000G and in a larger sample size. Another difference lies in the main aim of the UK10K project, which is to link phenotype/genotype relationships by using deeply phenotyped samples (from Twins UK and Avon Longitudinal Study of Parents and Children repositories) allowing a genetic comparison of shared traits.

These initiatives highlight a major advance in the field; to find and confirm disease causal mutations by sequencing thousands of individuals across multiple related diseases in order to gain statistical power required for accurate

sequence-based genotyping. The UK10K dataset provides a great control sample resource for ongoing smaller sequencing studies, something that was not available in 2009. The resource has also enabled imputation of low frequency (MAF <0.1%) variants into current GWAS data to increase power, and will be downloaded onto reference databases (e.g. dbSNP, RefSeq) providing deeper annotation of the genome. Along with the efforts from ENCODE, where features of genes have been significantly enhanced with high accuracy (Harrow, Frankish et al. 2012), these data will really benefit future studies aiming to locate genetic causal variants.

6.4 Where does 'missing heritability' of disease lie?

The main question the research in this thesis attempted to answer was: do rare variants carrying a high disease risk account for the missing heritability of disease? In light of recent findings in 42,000 subjects across six autoimmune diseases that found no rare variants at GWAS risk loci, showing the lack of support for synthetic and rare associations of large effect (Hunt, Mistry et al. 2013), similar conclusions can be made here. Although different approaches were taken (resequencing of GWAS risk loci and resequencing of candidate genes from an exome sequencing dataset), the aims of both studies were relatively similar and based on the CDRV hypothesis. If missing heritability does not lie in the rare allele mutational spectrum of disease, it disproves the CDRV hypothesis and other explanations should be sought. Proposals include epistasis, structural variants, parent-of-origin effects (Goriely and Wilkie 2010) and inherited epigenetic factors (Eichler, Flint et al. 2010).

To really know how much of the missing heritability to search for, it is useful to know how much is exactly missing and how much is unknown. Many reviews discuss flaws in estimates of heritability (Visscher, Hill et al. 2008; Tenesa and Haley 2013), particularly unaccounted for gene-gene and gene-environment interactions. Narrow sense heritability is more useful in predicting disease risk as it calculates the average disease risk passed on to offspring, but it only predicts contribution when there is no interaction between loci. Zuk et al

discuss ‘phantom heritability’ in their study, a phenomenon that can overestimate the proportion of phenotypic variation explained by additive effects of all variants inferred from population data by not including genetic interaction data, and assuming it is additive (Zuk, Hechter et al. 2012). An additive effect is defined as the risk allele acting in an independent and linear manner, with each independent allele having a cumulative effect. This in turn underestimates the overall heritability of a trait, initially predicted from pedigree-based studies. For example, the 71 GWAS-risk loci contributing to Crohn’s disease susceptibility only explains 23.2% of heritability (Franke, McGovern et al. 2010). This could possibly be explained by the use of tagging SNPs that can underestimate true risk from the causal allele, or by not accounting for multiple loci interactions. Epistasis is known to be of biological importance (Phillips 2008) and refers to interacting pairs of loci, where the phenotype of one locus is explained by the genotype at another (Carlborg and Haley 2004), and directly negates any additive allele effects. Estimation of epistatic effects can be incorporated into estimations of heritability, but will probably require extremely large sample sizes due to small individual interaction effects (Zuk, Hechter et al. 2012).

Additional undetected common variations of weak effects might account for missing heritability, as has been suggested by a study in yeast strains (Bloom, Ehrenreich et al. 2013). The authors attempted to explain differences between broad sense (phenotypic variation due to heritable factors) and narrow sense (phenotypic variation due to additive genetic factors) heritability by identifying the presence of two-locus interactions, however only small effects were detected. The findings suggest that missing narrow sense heritability is due to multiple weak effect loci, consistent with a study in human height, and any loci interactions only account for a small contribution (Yang, Benyamin et al. 2010). Statistical models have also been developed that take into account the contribution of SNPs just under the significance threshold for disease association, such as polygenic analysis which aggregates SNPs and tests their collective effect on phenotype (Purcell, Wray et al. 2009). Another study also applied a polygenic model to a simulated GWAS dataset and Bayesian

computation and supported the 'common causal variant weak effect size' contribution to heritability (Stahl, Wegmann et al. 2012). These studies collectively suggest that many low penetrant variants are yet to be discovered (Park, Wacholder et al. 2010).

6.5 Research update and concluding remarks

For rare-variant analysis in complex disease, new algorithms are consistently being developed. Only recently, a new method by Daniel MacArthur has been developed to process simultaneously a high number of exomes, resulting in cleaner data outputs and the discovery of new causal variants. Importantly, robust case control comparisons can be made (Vincent Plagnol, personal communication). Vincent Plagnol applied this algorithm to the 75-diseae case exome dataset, confirming many of the variants found using the calling algorithm applied in this thesis (SamTools), but some variants were missing or annotated differently. For example, one variant in *C4BPA* that was a false positive was not present in the new data, proving a cleaner output with the new algorithm, and a nonsynonymous variant in *C1QPB* was annotated as a synonymous SNP. Three nonsynonymous SNPs in *TRAF4, ACOT8* and *TNFRSF21*, respectively, were not present in the new dataset possibly because a different target region was used. So although the new algorithm has not drastically changed the results, the lack of nonsynonymous SNPs in *TRAF4, ACOT8* and *TNFRSF21* might have prevented them being chosen for resequencing, if the new calling method was applied at the time of analysis. Interestingly, none of these genes resulted in a gene-based test *P* value of <0.01. As this experiment was done many years ago, and significant improvements in statistical methods and experimental designs have been proposed, it might have been beneficial to exome sequence all affected individuals from coeliac pedigrees to locate the complete rare mutational spectrum of all affected individuals with a familial aggregation of disease.

In conclusion, the research shown in this thesis has provided extensive evidence about the contribution of rare variation in familial coeliac cases. The strategy of

sequencing multiply affected families, and deep follow up of candidate genes, has not identified new disease risk mutations. It could be that a subset of rare disease causing variants are only specific to one family and therefore will not be observed at the disease population level. Evidence from the BRK family, where all disease cases carry the novel C>T SNP in *TNFRSF21*, points towards this scenario. On the other hand, perhaps undetected common variants of weak effects (and other factors, e.g. environmental) may account for familial clustering of this common autoimmune disease.

Chapter 7

Future work and directions

This research in this thesis has attempted to target rare variation predisposing to CD. The current dataset leads to the conclusion that rare variation does not lead to disease risk in this family-based cohort, but future work based on the results here may possibly lead to a different outcome. Below is a list of proposals for a future PhD student:

- Test *EPAS1* in a larger case-control sample size: since the *P* value in the 4,608-sample dataset was 0.007 it may worth testing this gene in a larger sample size for any significant rare variant disease associations. A custom Taqman genotyping assay containing all *EPAS1* coding SNPs is the simplest experiment that will quickly answer this question.
- Resequence *CUBN*: this gene was too large to incorporate into the Fluidigm targeted resequencing assay. For this, a single-gene resequencing experiment in a medium case-control sample to begin with, (approximately 500 cases and controls) will test whether there is an excess of rare variation in coeliac cases compared to controls. Fluidigm resequencing technology can still be used here, but a 48-plex assay would suffice.

If one was to continue the search for rare variation in CD using familial samples to enrich for disease mutations and to account for familial clustering of disease, another design would be to exome sequence every affected individual from many coeliac pedigrees and compare to a matching population control dataset e.g. the UK10K exome dataset. This would provide a highly annotated dataset of every coding mutation in all sequenced individuals, however it may also lead to some data being discarded due to the sharing of chromosomal regions in families. Additionally, up to thousands of samples may be required to achieve the statistical power required in a complex disease, but in terms of sample design, many different approaches can be applied to achieve the best statistical result. It has been shown that exome sequencing trios and then performing a family-based association test may be particularly useful for rare variants, since the sample set would be robust to population stratification and Mendelian errors can be checked to reduce the false positive rate (De, Yip et al. 2013). Furthermore, there is evidence of increased sensitivity to find lower effect sizes

with the use of an enriched trio (one sibling from an ASP) in gene-based tests (Preston and Dudbridge 2013 *in press*). Since the study here utilized a family design and a case-control design on candidate genes, it provides a clue that the search for heritability may yield positive results if focused elsewhere. The following section discusses future research in the field of genetics that can be applied to CD, if one was to move away from attempting to locate rare disease variation.

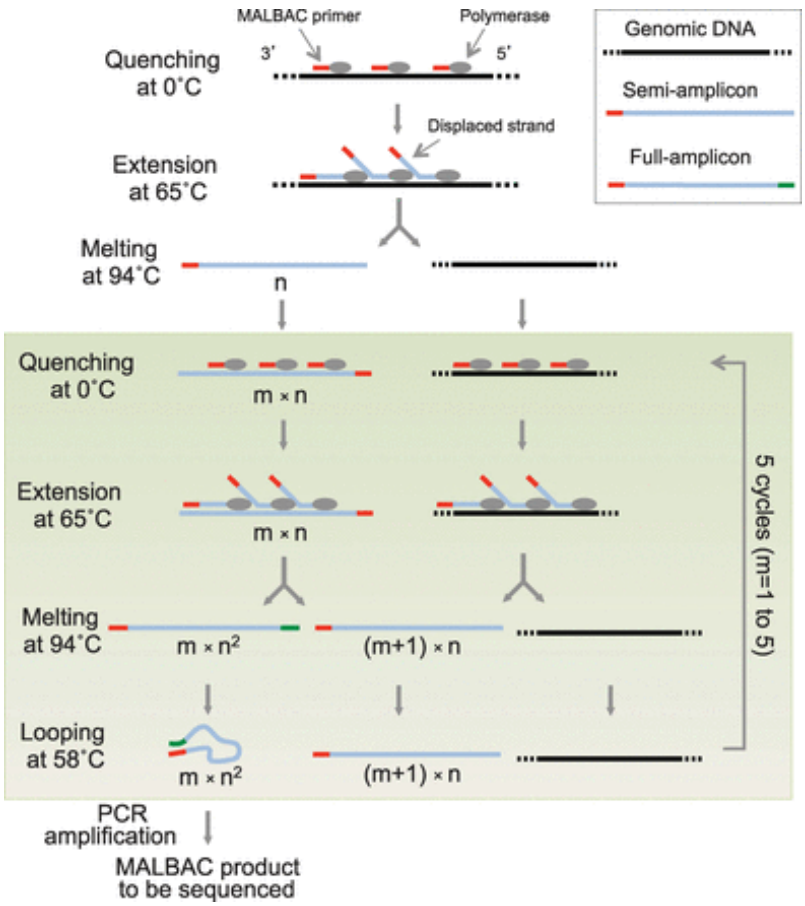
7.1 Further research in the field of coeliac disease genetics

Immunochip findings in CD show that most of the association signals are localized around transcription start sites and 3' UTR regions (Trynka, Hunt et al. 2011). Additionally, ENCODE findings revealed that most disease variants lie in regulatory regions and significant activity in these areas, including how much of the protein is produced rather than any modification to its structure, prove that there is much more occurring in non-coding regions than previously thought (Schaub, Boyle et al. 2012). For further genetic studies in CD, it may be a good idea to revisit findings from GWAS and fine mapping studies and attempt to link variant signals, even those not reaching GWAS significance as these probably fit under the umbrella of undetected loci, with a causal variant. Studies have shown that SNPs associated with common traits are enriched for expression quantitative trait loci (eQTL) (Lango Allen, Estrada et al. 2010; Nica, Montgomery et al. 2010; Nicolae, Gamazon et al. 2010), and even the last CD GWAS study found significant eQTLs in 20/38 non-HLA coeliac loci (Dubois, Trynka et al. 2010). The best example is the *SORT1* gene associated with plasma LDL concentration, where the associated variant modifies a CEBPB transcription factor binding site located in an enhancer, directly altering the expression of *SORT1* (Musunuru, Rader et al. 2010). Since common trait associated SNPs may be acting by altering gene regulatory regions, assessing cell subtypes with phenotypic associations might be able to identify true causal variations. The ENCODE project revealed SNPs associated with a disease phenotype were also associated with a specific cell type or transcription factor (Dunham, Kundaje et

al. 2012). A study by Trynka et al supports this finding in a study identifying chromatin marks in cell types (Trynka, Sandor et al. 2013). They show that chromatin peaks overlap with SNPs associated with common traits, e.g. 31 SNPs from RA regions overlap with chromatin marks in CD4+ regulatory T cells. Their findings highlight that cell type specific chromatin marks associated with phenotype can identify causal cell types. Looking deeper into immune cell subtypes in CD associated loci may therefore be the next step to further elucidate specific causal pathways.

Methods for single-cell analysis can be applied to enable deeper resolution of cell types. Methods published in the past have employed whole-genome amplification (WGA) of single cells (Zhang, Cui et al. 1992) and degenerate oligonucleotide PCR-based methods, but this technique generates short products not useful for many applications (Telenius, Carter et al. 1992). Multiple displacement amplification using hexamer primers and Phi 29 DNA polymerase generates much larger products (<10Kb) (Dean, Nelson et al. 2001) and is used for genotyping SNPs on Illumina chips, for example (Barker, Hansen et al. 2004). New methodologies are continuously being published to increase coverage required for single cell sequencing. A recent study reported a new WGA method named MALBAC, eliminating amplification bias associated with previous WGA methods (Zong, Lu et al. 2012). The authors designed primers to anneal randomly to single-cell DNA molecules, performed PCR with a DNA polymerase with displacement activity to create semi-amplicons, and then used these as templates to produce full amplicons (Figure 7.1). With this technique, they were able to identify SNVs from MALBAC-amplicons with no false positives and measure mutation rates of cancer cell lines.

Figure 7.1: MALBAC single-cell WGA to decrease amplification bias



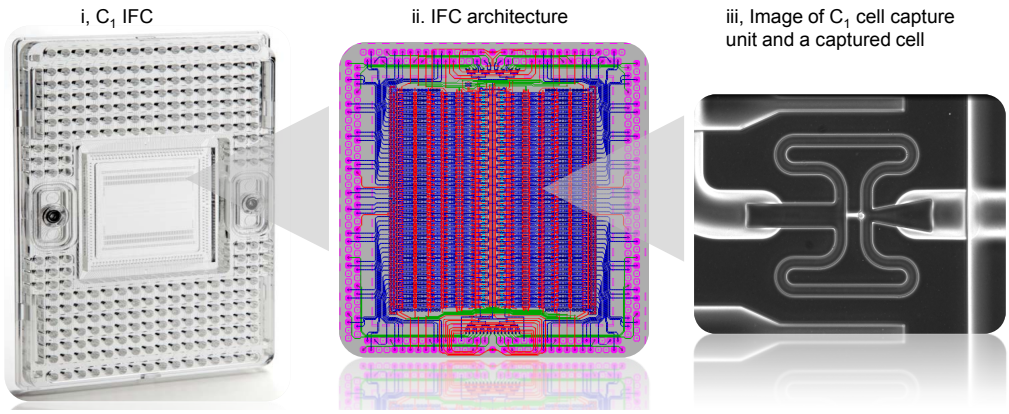
MALBAC = multiple annealing and looping-based amplification cycles. Taken from Zong, Lu et al. 2012.

Now, advances in NGS have enabled direct analysis of single cell genomes. A recently published study applied single-cell RNA sequencing in dendritic cells from bone marrows of mice to investigate heterogeneity in the response of these cells to lipopolysaccharide (Shalek, Satija et al. 2013). The study revealed interesting findings surrounding variation across single cells, such as bimodal splicing patterns with one isoform having a distinct function, differential activity in clusters of genes (i.e. in antiviral regulatory genes where co-variation in different cell transcripts helped to identify the antiviral cell circuit), and variation in expression patterns reflecting different cell developmental states. If

such variation is observed across immune cells, there is further scope in linking disease genotypes to single-cell phenotypes.

Commercial companies, such as Fluidigm, have also progressed onto single cell genomics. Fluidigm's intergrated microfluidics system has been developed for preparation of hundreds of cDNA libraries from single-cell samples for mRNA sequencing, enabling single-cell gene expression profiling. The technology combines 96 cDNA library preparations in parallel on an array (Figure 7.2). The amplified cDNA samples are then subjected to library preparation for Illumina sequencing. The method has shown to produce high quality sequencing libraries by Fluidigm's Research and Development group, and also confirmed transcriptional heterogeneity within homogenous cell populations (Shug, Chen et al. 2013). Using this technology to assess single-cell expression in CD might detect whether there are specific variations within cells from CD associated immune loci.

Figure 7.2: Fluidigm IFC cell capture illustration



The IFC array performs single–cell cDNA library preparations in tiny compartments. Taken from (Shug, Chen et al. 2013)

To summarize, the points outlined at the start of this chapter can be undertaken for further progression of locating rare variation in CD: *EPAS1* and *CUBN* might hold key genetic variants predisposing to CD risk and are likely candidate genes based on their function and findings in this thesis. If these experiments do not

prove fruitful, then there is the possibility that thousands of more individuals will require sequencing, which is costly, hence why collaborative initiatives such as UK10K and NHLBI Exome Sequencing Project have been set up. If one wants to move away from rare variant searching involving larger sample sizes, studies involving single cell genomics, with an aim of focusing on gene expression and specific gene interactions within cells and linking them with disease phenotype can provide a closer look at the functional consequences of mutations in certain cell types. This will develop a better understanding of genetic heterogeneity within cells and its relation to disease.

References

Abadie, V., L. M. Sollid, et al. (2011). "Integration of genetic and immunological insights into a model of celiac disease pathogenesis." Annu Rev Immunol **29**: 493-525.

Abecasis, G. R., D. Altshuler, et al. (2010). "A map of human genome variation from population-scale sequencing." Nature **467**(7319): 1061-1073.

Abecasis, G. R., A. Auton, et al. (2012). "An integrated map of genetic variation from 1,092 human genomes." Nature **491**(7422): 56-65.

Abecasis, G. R., S. S. Cherny, et al. (2002). "Merlin--rapid analysis of dense genetic maps using sparse gene flow trees." Nat Genet **30**(1): 97-101.

Abecasis, G. R., W. O. Cookson, et al. (2000). "Pedigree tests of transmission disequilibrium." Eur J Hum Genet **8**(7): 545-551.

Ahituv, N., N. Kavaslar, et al. (2007). "Medical sequencing at the extremes of human body mass." Am J Hum Genet **80**(4): 779-791.

Ahn, R., Y. C. Ding, et al. (2012). "Association analysis of the extended MHC region in celiac disease implicates multiple independent susceptibility loci." PLoS One **7**(5): e36926.

Al-toma, A., O. J. Visser, et al. (2007). "Autologous hematopoietic stem cell transplantation in refractory celiac disease with aberrant T cells." Blood **109**(5): 2243-2249.

Albrechtsen, A., N. Grarup, et al. (2013). "Exome sequencing-driven discovery of coding polymorphisms associated with common metabolic phenotypes." Diabetologia **56**(2): 298-310.

Alkan, C., J. M. Kidd, et al. (2009). "Personalized copy number and segmental duplication maps using next-generation sequencing." Nat Genet **41**(10): 1061-1067.

Aminoff, M., J. E. Carter, et al. (1999). "Mutations in CUBN, encoding the intrinsic factor-vitamin B12 receptor, cubilin, cause hereditary megaloblastic anaemia 1." Nat Genet **21**(3): 309-313.

Amundsen, S. S., A. T. Naluai, et al. (2004). "Genetic analysis of the CD28/CTLA4/ICOS (CELIAC3) region in coeliac disease." Tissue Antigens **64**(5): 593-599.

Amundsen, S. S., J. Rundberg, et al. (2010). "Four novel coeliac disease regions replicated in an association study of a Swedish-Norwegian family cohort." Genes Immun **11**(1): 79-86.

Arakawa, S., A. Takahashi, et al. (2011). "Genome-wide association study identifies two susceptibility loci for exudative age-related macular degeneration in the Japanese population." Nat Genet **43**(10): 1001-1004.

Asimit, J. and E. Zeggini (2010). "Rare variant association analysis methods for complex traits." Annu Rev Genet **44**: 293-308.

Babron, M. C., S. Nilsson, et al. (2003). "Meta and pooled analysis of European coeliac disease data." Eur J Hum Genet **11**(11): 828-834.

Bahram, S. (2000). "MIC genes: from genetics to biology." Adv Immunol **76**: 1-60.

Bai, D., P. Brar, et al. (2005). "Effect of gender on the manifestations of celiac disease: evidence for greater malabsorption in men." Scand J Gastroenterol **40**(2): 183-187.

Bansal, V., O. Libiger, et al. (2010). "Statistical analysis strategies for association studies involving rare variants." Nat Rev Genet **11**(11): 773-785.

Baranzini, S. E. (2009). "The genetics of autoimmune diseases: a networked perspective." Curr Opin Immunol **21**(6): 596-605.

Barker, D. L., M. S. Hansen, et al. (2004). "Two methods of whole-genome amplification enable accurate genotyping across a 2320-SNP linkage panel." Genome Res **14**(5): 901-907.

Barrett, J. C., S. Hansoul, et al. (2008). "Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease." Nat Genet **40**(8): 955-962.

Bensellam, M., B. Duvillie, et al. (2012). "Glucose-induced O(2) consumption activates hypoxia inducible factors 1 and 2 in rat insulin-secreting pancreatic beta-cells." PLoS One **7**(1): e29807.

Bertin, J., L. Wang, et al. (2001). "CARD11 and CARD14 are novel caspase recruitment domain (CARD)/membrane-associated guanylate kinase (MAGUK) family members that interact with BCL10 and activate NF-kappa B." J Biol Chem **276**(15): 11877-11882.

Bhatia, G., V. Bansal, et al. (2010). "A covering method for detecting genetic associations between rare variants and common phenotypes." PLoS Comput Biol **6**(10): e1000954.

Bilguvar, K., A. K. Ozturk, et al. (2010). "Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations." Nature **467**(7312): 207-210.

Blekhman, R., O. Man, et al. (2008). "Natural selection on genes that underlie human disease susceptibility." Curr Biol **18**(12): 883-889.

Bloom, J. S., I. M. Ehrenreich, et al. (2013). "Finding the sources of missing heritability in a yeast cross." Nature **494**(7436): 234-237.

Bodmer, W. and C. Bonilla (2008). "Common and rare variants in multifactorial susceptibility to common diseases." Nat Genet **40**(6): 695-701.

Boger, C. A., M. H. Chen, et al. (2011). "CUBN is a gene locus for albuminuria." J Am Soc Nephrol **22**(3): 555-570.

Boileau, C., D. C. Guo, et al. (2012). "TGFB2 mutations cause familial thoracic aortic aneurysms and dissections associated with mild systemic features of Marfan syndrome." Nat Genet **44**(8): 916-921.

Borecki, I. B. and M. A. Province (2008). "Genetic and Genomic Discovery Using Family Studies " Circulation **118**: 1057-1063.

Botstein, D. and N. Risch (2003). "Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease." Nat Genet **33 Suppl**: 228-237.

Botstein, D., R. L. White, et al. (1980). "Construction of a genetic linkage map in man using restriction fragment length polymorphisms." Am J Hum Genet **32**(3): 314-331.

Bowden, D. W., S. S. An, et al. (2010). "Molecular basis of a linkage peak: exome sequencing and family-based analysis identify a rare genetic variant in the ADIPOQ gene in the IRAS Family Study." Hum Mol Genet **19**(20): 4112-4120.

Boyko, A. R., S. H. Williamson, et al. (2008). "Assessing the evolutionary impact of amino acid mutations in the human genome." PLoS Genet **4**(5): e1000083.

Bravo, H. C. and R. A. Irizarry (2010). "Model-based quality assessment and base-calling for second-generation sequencing data." Biometrics **66**(3): 665-674.

Brett, P. M., J. Y. Yiannakou, et al. (1999). "Common HLA alleles, rather than rare mutants, confer susceptibility to coeliac disease." Ann Hum Genet **63**(Pt 3): 217-225.

Brophy, K., A. W. Ryan, et al. (2006). "Haplotypes in the CTLA4 region are associated with coeliac disease in the Irish population." Genes Immun **7**(1): 19-26.

Brown, W. M., J. Pierce, et al. (2009). "Overview of the MHC fine mapping data." Diabetes Obesity & Metabolism **11 Suppl 1**: 2-7.

Burton, P. R., D. G. Clayton, et al. (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." Nature **447**(7145): 661-678.

Bustamante, C. D., A. Fledel-Alon, et al. (2005). "Natural selection on protein-coding genes in the human genome." Nature **437**(7062): 1153-1157.

Byun, M., A. Abhyankar, et al. (2010). "Whole-exome sequencing-based discovery of STIM1 deficiency in a child with fatal classic Kaposi sarcoma." J Exp Med **207**(11): 2307-2312.

Caliskan, M., J. X. Chong, et al. (2011). "Exome Sequencing Reveals a Novel Mutation for Autosomal Recessive Nonsyndromic Mental Retardation in the TECR Gene on Chromosome 19p13." Hum Mol Genet.

Cardenas-Roldan, J., A. Rojas-Villarraga, et al. (2013). "How do autoimmune diseases cluster in families? A systematic review and meta-analysis." BMC Med **11**: 73.

Cargill, M., D. Altshuler, et al. (1999). "Characterization of single-nucleotide polymorphisms in coding regions of human genes." Nat Genet **22**(3): 231-238.

Carlborg, O. and C. S. Haley (2004). "Epistasis: too often neglected in complex trait studies?" Nat Rev Genet **5**(8): 618-625.

Carlson, C. S., M. A. Eberle, et al. (2004). "Mapping complex disease loci in whole-genome association studies." Nature **429**(6990): 446-452.

Casbon, J. A., R. J. Osborne, et al. (2011). "A method for counting PCR template molecules with application to next-generation sequencing." Nucleic Acids Res **39**(12): e81.

Castellanos-Rubio, A., I. Santin, et al. (2009). "TH17 (and TH1) signatures of intestinal biopsies of CD patients in response to gliadin." Autoimmunity **42**(1): 69-73.

Catassi, C., I. M. Ratsch, et al. (1999). "Why is coeliac disease endemic in the people of the Sahara?" Lancet **354**(9179): 647-648.

Cepek, K. L., C. M. Parker, et al. (1993). "Integrin alpha E beta 7 mediates adhesion of T lymphocytes to epithelial cells." J Immunol **150**(8 Pt 1): 3459-3470.

Chang, F., U. Mahadeva, et al. (2005). "Pathological and clinical significance of increased intraepithelial lymphocytes (IELs) in small bowel mucosa." APMIS **113**(6): 385-399.

Chen, R., E. A. Stahl, et al. (2011). "Fine mapping the TAGAP risk locus in rheumatoid arthritis." Genes Immun.

Cirulli, E. T. and D. B. Goldstein (2010). "Uncovering the roles of rare variants in common disease through whole-genome sequencing." Nature Reviews Genetics **11**(6): 415-425.

Clancy, R. M., M. C. Marion, et al. (2010). "Identification of candidate loci at 6p21 and 21q22 in a genome-wide association study of cardiac manifestations of neonatal lupus." Arthritis Rheum **62**(11): 3415-3424.

Clot, F., M. C. Fulchignoni-Lataud, et al. (1999). "Linkage and association study of the CTLA-4 region in coeliac disease for Italian and Tunisian populations." Tissue Antigens **54**(5): 527-530.

Cohen, J. C., R. S. Kiss, et al. (2004). "Multiple rare alleles contribute to low plasma levels of HDL cholesterol." Science **305**(5685): 869-872.

Consortium, T. I. H. (2003). "The International HapMap Project." Nature **426**(6968): 789-796.

Cooper, D. N., E. V. Ball, et al. (1998). "The human gene mutation database." Nucleic Acids Res **26**(1): 285-287.

Cooper, J. D., M. J. Simmonds, et al. (2012). "Seven newly identified loci for autoimmune thyroid disease." Hum Mol Genet **21**(23): 5202-5208.

Cooper, J. D., D. J. Smyth, et al. (2008). "Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci." Nat Genet **40**(12): 1399-1401.

Cortes, A. and M. A. Brown (2011). "Promise and pitfalls of the Immunochip." Arthritis Res Ther **13**(1): 101.

Couch, F. J., X. Wang, et al. (2013). "Genome-wide association study in BRCA1 mutation carriers identifies novel loci associated with breast and ovarian cancer risk." PLoS Genet **9**(3): e1003212.

Coventry, A., L. M. Bull-Otterson, et al. (2010). "Deep resequencing reveals excess rare recent variants consistent with explosive population growth." Nat Commun **1**: 131.

Croese, J., S. T. Gaze, et al. (2013). "Changed gluten immunity in celiac disease by Necator americanus provides new insights into autoimmunity." Int J Parasitol.

Cummins, A. G. and I. C. Roberts-Thomson (2009). "Prevalence of celiac disease in the Asia-Pacific region." J Gastroenterol Hepatol **24**(8): 1347-1351.

Curtis, D. and P. C. Sham (1995). "Model-free linkage analysis using likelihoods." Am J Hum Genet **57**(3): 703-716.

de Bakker, P. I., M. A. Ferreira, et al. (2008). "Practical aspects of imputation-driven meta-analysis of genome-wide association studies." Hum Mol Genet **17**(R2): R122-128.

de Bakker, P. I., G. McVean, et al. (2006). "A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC." Nat Genet **38**(10): 1166-1172.

De, G., W. K. Yip, et al. (2013). "Rare variant analysis for family-based design." PLoS One **8**(1): e48495.

Dean, F. B., J. R. Nelson, et al. (2001). "Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification." Genome Res **11**(6): 1095-1099.

Di Sabatino, A. and G. R. Corazza (2009). "Coeliac disease." Lancet **373**(9673): 1480-1493.

Dicke, W. K., H. A. W. HA, et al. (1953). "Coeliac disease. II. The presence in wheat of a factor having a deleterious effect in cases of coeliac disease
." Acta Paediatr **42**: 34-42.

Dieterich, W., T. Ehnis, et al. (1997). "Identification of tissue transglutaminase as the autoantigen of celiac disease." Nature Medicine **3**(7): 797-801.

Diogo, D., F. Kurreeman, et al. (2013). "Rare, low-frequency, and common variants in the protein-coding sequence of biological candidate genes from GWASs contribute to risk of rheumatoid arthritis." Am J Hum Genet **92**(1): 15-27.

Djilali-Saiah, I., J. Schmitz, et al. (1998). "CTLA-4 gene polymorphism is associated with predisposition to coeliac disease." Gut **43**(2): 187-189.

Do, R., S. Kathiresan, et al. (2012). "Exome sequencing and complex disease: practical aspects of rare variant association studies." Hum Mol Genet **21**(R1): R1-9.

Doyle, G. A., J. P. Dahl, et al. (2011). "Association study of polymorphisms in the autosomal mitochondrial complex I subunit gene, NADH dehydrogenase (ubiquinone) flavoprotein 2, and bipolar disorder." Psychiatr Genet **21**(1): 51-52.

Dube, C., A. Rostom, et al. (2005). "The prevalence of celiac disease in average-risk and at-risk Western European populations: a systematic review." Gastroenterology **128**(4 Suppl 1): S57-67.

Dubois, P. C., G. Trynka, et al. (2010). "Multiple common variants for celiac disease influencing immune gene expression." Nat Genet **42**(4): 295-302.

Dubois, P. C. and D. A. van Heel (2008). "Translational mini-review series on the immunogenetics of gut disease: immunogenetics of coeliac disease." Clin Exp Immunol **153**(2): 162-173.

Dunham, I., A. Kundaje, et al. (2012). "An integrated encyclopedia of DNA elements in the human genome." Nature **489**(7414): 57-74.

Durbin, R. M., G. R. Abecasis, et al. (2010). "A map of human genome variation from population-scale sequencing." Nature **467**(7319): 1061-1073.

Economou, M., T. A. Trikalinos, et al. (2004). "Differential effects of NOD2 variants on Crohn's disease risk and phenotype in diverse populations: a metaanalysis." Am J Gastroenterol **99**(12): 2393-2404.

Eichler, E. E., J. Flint, et al. (2010). "Missing heritability and strategies for finding the underlying causes of complex disease." Nat Rev Genet **11**(6): 446-450.

Eller, E., P. Vardi, et al. (2006). "Celiac disease and HLA in a Bedouin kindred." Hum Immunol **67**(11): 940-950.

Emond, M. J., T. Louie, et al. (2012). "Exome sequencing of extreme phenotypes identifies DCTN4 as a modifier of chronic Pseudomonas aeruginosa infection in cystic fibrosis." Nat Genet **44**(8): 886-889.

ESPGHAN (1990). "Revised criteria for diagnosis of coeliac disease. Report of Working Group of European Society of Paediatric Gastroenterology and Nutrition." Arch Dis Child **65**(8): 909-911.

Eyre, S., J. Bowes, et al. (2012). "High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis." Nat Genet **44**(12): 1336-1340.

Fay, J. C., G. J. Wyckoff, et al. (2001). "Positive and negative selection on the human genome." Genetics **158**(3): 1227-1234.

Fearnhead, N. S., J. L. Wilding, et al. (2004). "Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas." Proc Natl Acad Sci U S A **101**(45): 15992-15997.

Ferguson, A., E. Arranz, et al. (1993). "Clinical and pathological spectrum of coeliac disease--active, silent, latent, potential." Gut **34**(2): 150-151.

Fernandez, S., I. J. Molina, et al. (2011). "Characterization of gliadin-specific Th17 cells from the mucosa of celiac disease patients." Am J Gastroenterol **106**(3): 528-538.

Festen, E. A., P. Goyette, et al. (2011). "A meta-analysis of genome-wide association scans identifies IL18RAP, PTPN2, TAGAP, and PUS10 as shared risk loci for Crohn's disease and celiac disease." PLoS Genet **7**(1): e1001283.

Fieuw, A., B. De Wilde, et al. (2012). "Cancer gene prioritization for targeted resequencing using FitSNP scores." PLoS One **7**(3): e31333.

Fodinger, M., O. F. Wagner, et al. (2001). "Recent insights into the molecular genetics of the homocysteine metabolism." Kidney Int Suppl **78**: S238-242.

Franke, A., D. P. McGovern, et al. (2010). "Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci." Nat Genet **42**(12): 1118-1125.

Fu, W., T. D. O'Connor, et al. (2013). "Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants." Nature **493**(7431): 216-220.

Gao, X., T. Haritunians, et al. (2012). "Genotype Imputation for Latinos Using the HapMap and 1000 Genomes Project Reference Panels." Front Genet **3**: 117.

Garner, C. P., J. A. Murray, et al. (2009). "Replication of celiac disease UK genome-wide association study results in a US population." Hum Mol Genet **18**(21): 4219-4225.

Gibson, G. (2011). "Rare and common variants: twenty arguments." Nat Rev Genet **13**(2): 135-145.

Glas, J., J. Stallhofer, et al. (2009). "Novel genetic risk markers for ulcerative colitis in the IL2/IL21 region are in epistasis with IL23R and suggest a common genetic background for ulcerative colitis and celiac disease." Am J Gastroenterol **104**(7): 1737-1744.

Glazov, E. A., A. Zankl, et al. (2011). "Whole-Exome Re-Sequencing in a Family Quartet Identifies POP1 Mutations As the Cause of a Novel Skeletal Dysplasia." PLoS Genet **7**(3): e1002027.

Golan, D. and S. Rosset (2011). "Accurate estimation of heritability in genome wide studies using random effects models." Bioinformatics **27**(13): i317-323.

Goriely, A. and A. O. Wilkie (2010). "Missing heritability: paternal age effect mutations and selfish spermatogonia." Nat Rev Genet **11**(8): 589.

Gravel, S., B. M. Henn, et al. (2011). "Demographic history and rare allele sharing among human populations." Proc Natl Acad Sci U S A **108**(29): 11983-11988.

Greco, L., M. C. Babron, et al. (2001). "Existence of a genetic risk factor on chromosome 5q in Italian coeliac disease families." Ann Hum Genet **65**(Pt 1): 35-41.

Greco, L., G. Corazza, et al. (1998). "Genome search in celiac disease." Am J Hum Genet **62**(3): 669-675.

Greco, L., R. Romino, et al. (2002). "The first large population based twin study of coeliac disease." Gut **50**(5): 624-628.

Gutierrez-Achury, J., R. C. de Almeida, et al. (2011). "Shared genetics in celiac disease and other immune-mediated diseases." J Intern Med.

Haines, J. L., M. A. Hauser, et al. (2005). "Complement factor H variant increases the risk of age-related macular degeneration." Science **308**(5720): 419-421.

Han, F. and W. Pan (2010). "A data-adaptive sum test for disease association with multiple common or rare variants." Hum Hered **70**(1): 42-54.

Hansson, T., A. K. Ulfgren, et al. (2002). "Transforming growth factor-beta (TGF-beta) and tissue transglutaminase expression in the small intestine in children with coeliac disease." Scand J Immunol **56**(5): 530-537.

Harrow, J., A. Frankish, et al. (2012). "GENCODE: the reference human genome annotation for The ENCODE Project." Genome Res **22**(9): 1760-1774.

Hattori, N., H. Yoshino, et al. (1998). "Genotype in the 24-kDa subunit gene (NDUFV2) of mitochondrial complex I and susceptibility to Parkinson disease." Genomics **49**(1): 52-58.

He, C., S. Hamon, et al. (2009). "MHC fine mapping of human type 1 diabetes using the T1DGC data." Diabetes Obesity & Metabolism **11**: 53-59.

Heap, G. A. and D. A. van Heel (2009). "Genetics and pathogenesis of coeliac disease." Semin Immunol **21**(6): 346-354.

Hedges, D. J., T. Guettouche, et al. (2011). "Comparison of three targeted enrichment strategies on the SOLiD sequencing platform." PLoS One **6**(4): e18595.

Heinzen, E. L., C. Depondt, et al. (2012). "Exome sequencing followed by large-scale genotyping fails to identify single rare variants of large effect in idiopathic generalized epilepsy." Am J Hum Genet **91**(2): 293-302.

Helbig, I., S. E. Hodge, et al. (2013). "Familial cosegregation of rare genetic variants with disease in complex disorders." Eur J Hum Genet **21**(4): 444-450.

Helbig, I., H. C. Mefford, et al. (2009). "15q13.3 microdeletions increase risk of idiopathic generalized epilepsy." Nat Genet **41**(2): 160-162.
Hirschhorn, J. N., K. Lohmueller, et al. (2002). "A comprehensive review of genetic association studies." Genet Med **4**(2): 45-61.
Hodgkinson, A. and A. Eyre-Walker (2010). "Human triallelic sites: evidence for a new mutational mechanism?" Genetics **184**(1): 233-241.
Hoischen, A., B. W. van Bon, et al. (2010). "De novo mutations of SETBP1 cause Schinzel-Giedion syndrome." Nat Genet **42**(6): 483-485.
Holm, H., D. F. Gudbjartsson, et al. (2011). "A rare variant in MYH6 is associated with high risk of sick sinus syndrome." Nat Genet **43**(4): 316-320.
Holtmann, M. H. and M. F. Neurath (2004). "T helper cell polarisation in coeliac disease: any (T-)bet ?" Gut **53**(8): 1065-1067.
Homer, C. R., A. L. Richmond, et al. (2010). "ATG16L1 and NOD2 interact in an autophagy-dependent antibacterial pathway implicated in Crohn's disease pathogenesis." Gastroenterology **139**(5): 1630-1641, 1641 e1631-1632.
Horton, R., L. Wilming, et al. (2004). "Gene map of the extended human MHC." Nat Rev Genet **5**(12): 889-899.
Howald, C., A. Tanzer, et al. (2012). "Combining RT-PCR-seq and RNA-seq to catalog all genic elements encoded in the human genome." Genome Res **22**(9): 1698-1710.
Howson, J. M., N. M. Walker, et al. (2009). "Confirmation of HLA class II independent type 1 diabetes associations in the major histocompatibility complex including HLA-B and HLA-A." Diabetes Obesity & Metabolism **11 Suppl 1**: 31-45.
Huang, J., D. Ellinghaus, et al. (2012). "1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data." Eur J Hum Genet **20**(7): 801-805.
Hue, S., J. J. Mention, et al. (2004). "A direct role for NKG2D/MICA interaction in villous atrophy during celiac disease." Immunity **21**(3): 367-377.
Huebner, C., I. Petermann, et al. (2007). "Triallelic single nucleotide polymorphisms and genotyping error in genetic epidemiology studies: MDR1 (ABCB1) G2677/T/A as an example." Cancer Epidemiol Biomarkers Prev **16**(6): 1185-1192.
Hughes, T., X. Kim-Howard1, et al. (2011). "Fine mapping and trans-ethnic genotyping establish IL2/IL21 genetic association with lupus and localize this genetic effect to IL21." Arthritis & Rheumatism.
Hugot, J. P., M. Chamaillard, et al. (2001). "Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease." Nature **411**(6837): 599-603.
Hunt, K. A., D. P. McGovern, et al. (2005). "A common CTLA4 haplotype associated with coeliac disease." Eur J Hum Genet **13**(4): 440-444.
Hunt, K. A., V. Mistry, et al. (2013). "Negligible impact of rare autoimmune-locus coding-region variants on missing heritability." Nature.
Hunt, K. A., D. J. Smyth, et al. (2012). "Rare and functional SIAE variants are not associated with autoimmune disease risk in up to 66,924 individuals of European ancestry." Nat Genet **44**(1): 3-5.

Hunt, K. A., A. Zhernakova, et al. (2008). "Newly identified genetic risk variants for celiac disease related to the immune response." Nat Genet **40**(4): 395-402.

Ingram, V. M. (1957). "Gene mutations in human hemoglobin: the chemical difference between normal and sickle cell hemoglobin." Nature **180**: 326-328.

Ionita-Laza, I., J. D. Buxbaum, et al. (2011). "A new testing strategy to identify rare variants with either risk or protective effect on disease." PLoS Genet **7**(2): e1001289.

Ionita-Laza, I. and R. Ottman (2011). "Study designs for identification of rare disease variants in complex diseases: the utility of family-based designs." Genetics **189**(3): 1061-1068.

Ji, W., J. N. Foo, et al. (2008). "Rare independent mutations in renal salt handling genes contribute to blood pressure variation." Nat Genet **40**(5): 592-599.

Jin, Y., S. A. Birlea, et al. (2010). "Variant of TYR and autoimmunity susceptibility loci in generalized vitiligo." N Engl J Med **362**(18): 1686-1697.

Johansen, C. T., J. Wang, et al. (2010). "Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia." Nat Genet **42**(8): 684-687.

Johnson, J. O., J. Mandrioli, et al. (2010). "Exome sequencing reveals VCP mutations as a cause of familial ALS." Neuron **68**(5): 857-864.

Jordan, C. T., L. Cao, et al. (2012). "Rare and common variants in CARD14, encoding an epidermal regulator of NF-kappaB, in psoriasis." Am J Hum Genet **90**(5): 796-808.

Jordan, C. T., L. Cao, et al. (2012). "PSORS2 is due to mutations in CARD14." Am J Hum Genet **90**(5): 784-795.

Jostins, L., S. Ripke, et al. (2012). "Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease." Nature **491**(7422): 119-124.

Jungel, A., J. H. Distler, et al. (2004). "Expression of interleukin-21 receptor, but not interleukin-21, in synovial fibroblasts and synovial macrophages of patients with rheumatoid arthritis." Arthritis Rheum **50**(5): 1468-1476.

Junker, Y., S. Zeissig, et al. (2012). "Wheat amylase trypsin inhibitors drive intestinal inflammation via activation of toll-like receptor 4." J Exp Med **209**(13): 2395-2408.

Juran, B. D., G. M. Hirschfield, et al. (2012). "Immunochip analyses identify a novel risk locus for primary biliary cirrhosis at 13q14, multiple independent associations at four established risk loci and epistasis between 1p31 and 7q32 risk variants." Hum Mol Genet **21**(23): 5209-5221.

Karell, K., A. S. Louka, et al. (2003). "HLA types in celiac disease patients not carrying the DQA1*05-DQB1*02 (DQ2) heterodimer: results from the European Genetics Cluster on Celiac Disease." Hum Immunol **64**(4): 469-477.

Kaukinen, K., J. Partanen, et al. (2002). "HLA-DQ typing in the diagnosis of celiac disease." Am J Gastroenterol **97**(3): 695-699.

Kazma, R. and J. N. Bailey (2011). "Population-based and family-based designs to analyze rare variants in complex diseases." Genet Epidemiol **35 Suppl 1**: S41-47.

Keller, B. J., S. Martini, et al. (2012). "Linking variants from genome-wide association analysis to function via transcriptional network analysis." Methods Mol Biol **910**: 297-308.

Kerem, B., J. M. Rommens, et al. (1989). "Identification of the cystic fibrosis gene: genetic analysis." Science **245**(4922): 1073-1080.

Kiezun, A., K. Garimella, et al. (2012). "Exome sequencing and the genetic basis of complex traits." Nat Genet **44**(6): 623-630.

Kiezun, A., S. L. Pulit, et al. (2013). "Deleterious alleles in the human genome are on average younger than neutral alleles of the same frequency." PLoS Genet **9**(2): e1003301.

King, A. L., J. S. Fraser, et al. (2001). "Coeliac disease: follow-up linkage study provides further support for existence of a susceptibility locus on chromosome 11p11." Ann Hum Genet **65**(Pt 4): 377-386.

King, A. L., S. J. Moodie, et al. (2003). "Coeliac disease: investigation of proposed causal variants in the CTLA4 gene region." Eur J Immunogenet **30**(6): 427-432.

King, A. L., S. J. Moodie, et al. (2002). "CTLA-4/CD28 gene region is associated with genetic susceptibility to coeliac disease in UK families." J Med Genet **39**(1): 51-54.

King, A. L., J. Y. Yiannakou, et al. (2000). "A genome-wide family-based linkage study of coeliac disease." Ann Hum Genet **64**(Pt 6): 479-490.

King, C. R., P. J. Rathouz, et al. (2010). "An evolutionary framework for association testing in resequencing studies." PLoS Genet **6**(11): e1001202.

Kingsmore, S. F., D. L. Dinwiddie, et al. (2011). "Adopting orphans: comprehensive genetic testing of Mendelian diseases of childhood by next-generation sequencing." Expert Rev Mol Diagn **11**(8): 855-868.

Kong, A. and N. J. Cox (1997). "Allele-sharing models: LOD scores and accurate linkage tests." Am J Hum Genet **61**(5): 1179-1188.

Kong, Y. (2011). "Btrim: a fast, lightweight adapter and quality trimming program for next-generation sequencing technologies." Genomics **98**(2): 152-153.

Kotowski, I. K., A. Pertsemlidis, et al. (2006). "A spectrum of PCSK9 alleles contributes to plasma levels of low-density lipoprotein cholesterol." Am J Hum Genet **78**(3): 410-422.

Kotze, L. M. (2009). "Celiac disease in Brazilian patients: associations, complications and causes of death. Forty years of clinical experience." Arq Gastroenterol **46**(4): 261-269.

Kozarewa, I., Z. Ning, et al. (2009). "Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes." Nat Methods **6**(4): 291-295.

Kozarewa, I. and D. J. Turner (2011). "Amplification-free library preparation for paired-end Illumina sequencing." Methods Mol Biol **733**: 257-266.

Krawitz, P. M., M. R. Schweiger, et al. (2010). "Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome." Nat Genet **42**(10): 827-829.

Kruglyak, L. (2008). "The road to genome-wide association studies." Nat Rev Genet **9**(4): 314-318.

Kruglyak, L., M. J. Daly, et al. (1996). "Parametric and nonparametric linkage analysis: a unified multipoint approach." Am J Hum Genet **58**(6): 1347-1363.

Krumm, N., P. H. Sudmant, et al. (2012). "Copy number variation detection and genotyping from exome sequence data." Genome Res **22**(8): 1525-1532.

Kryukov, G. V., L. A. Pennacchio, et al. (2007). "Most rare missense alleles are deleterious in humans: implications for complex disease and association studies." Am J Hum Genet **80**(4): 727-739.

Kryukov, G. V., A. Shpunt, et al. (2009). "Power of deep, all-exon resequencing for discovery of human trait genes." Proc Natl Acad Sci U S A **106**(10): 3871-3876.

Kumar, R., R. Goswami, et al. (2007). "Association and interaction of the TNF-alpha gene with other pro- and anti-inflammatory cytokine genes and HLA genes in patients with type 1 diabetes from North India." Tissue Antigens **69**(6): 557-567.

Kwiatkowski, D., A. V. Hill, et al. (1990). "TNF concentration in fatal cerebral, non-fatal cerebral, and uncomplicated Plasmodium falciparum malaria." Lancet **336**(8725): 1201-1204.

Lander, E. S. (1996). "The new genomics: global views of biology." Science **274**(5287): 536-539.

Lander, E. S. and D. Botstein (1986). "Strategies for studying heterogeneous genetic traits in humans by using a linkage map of restriction fragment length polymorphisms." Proc Natl Acad Sci U S A **83**(19): 7353-7357.

Lango Allen, H., K. Estrada, et al. (2010). "Hundreds of variants clustered in genomic loci and biological pathways affect human height." Nature **467**(7317): 832-838.

Lanzini, A., F. Lanzarotto, et al. (2009). "Complete recovery of intestinal mucosa occurs very rarely in adult coeliac patients despite adherence to gluten-free diet." Aliment Pharmacol Ther **29**(12): 1299-1308.

Lee, S. H., N. R. Wray, et al. (2011). "Estimating missing heritability for disease from genome-wide association studies." Am J Hum Genet **88**(3): 294-305.

Lee-Kirsch, M. A., M. Gong, et al. (2007). "Mutations in the gene encoding the 3'-5' DNA exonuclease TREX1 are associated with systemic lupus erythematosus." Nat Genet **39**(9): 1065-1067.

Lee-Kirsch, M. A., M. Gong, et al. (2006). "Familial chilblain lupus, a monogenic form of cutaneous lupus erythematosus, maps to chromosome 3p." Am J Hum Genet **79**(4): 731-737.

Lehne, B., C. M. Lewis, et al. (2011). "Exome localization of complex disease association signals." BMC Genomics **12**: 92.

Lesage, S., H. Zouali, et al. (2002). "CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease." Am J Hum Genet **70**(4): 845-857.

Lescai, F. and C. Franceschi (2010). "The impact of phenocopy on the genetic analysis of complex traits." PLoS One **5**(7): e11876.

Li, B. and S. M. Leal (2008). "Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data." Am J Hum Genet **83**(3): 311-321.

Li, D., J. P. Lewinger, et al. (2011). "Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies." Genet Epidemiol **35**(8): 790-799.

Li, M. X., H. S. Gui, et al. (2012). "A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases." Nucleic Acids Res **40**(7): e53.

Li, Y., N. Vinckenbosch, et al. (2010). "Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants." Nat Genet **42**(11): 969-972.

Liao, M., F. Ye, et al. (2012). "Genome-wide association study identifies common variants at TNFRSF13B associated with IgG level in a healthy Chinese male population." Genes Immun **13**(6): 509-513.

Lim, E. T., S. Raychaudhuri, et al. (2013). "Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders." Neuron **77**(2): 235-242.

Lin, Z., Q. Chen, et al. (2012). "Exome sequencing reveals mutations in TRPV3 as a cause of Olmsted syndrome." Am J Hum Genet **90**(3): 558-564.

Liu, D. J. and S. M. Leal (2010). "A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions." PLoS Genet **6**(10): e1001156.

Liu, D. J. and S. M. Leal (2012). "Estimating genetic effects and quantifying missing heritability explained by identified rare-variant associations." Am J Hum Genet **91**(4): 585-596.

Liu, J., S. H. Juo, et al. (2002). "Genomewide linkage analysis of celiac disease in Finnish families." Am J Hum Genet **70**(1): 51-59.

Liu, J., S. Na, et al. (2001). "Enhanced CD4+ T cell proliferation and Th2 cytokine production in DR6-deficient mice." Immunity **15**(1): 23-34.

Liu, J. Z., M. A. Almarri, et al. (2012). "Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis." Nat Genet **44**(10): 1137-1141.

Lohi, S., K. Mustalahti, et al. (2007). "Increasing prevalence of coeliac disease over time." Aliment Pharmacol Ther **26**(9): 1217-1225.

Losowsky, M. S. (2008). "A history of coeliac disease." Dig Dis **26**(2): 112-120.

Louis-Dit-Picard, H., J. Barc, et al. (2012). "KLHL3 mutations cause familial hyperkalemic hypertension by impairing ion transport in the distal nephron." Nat Genet **44**(4): 456-460, S451-453.

Lowe, C. E., J. D. Cooper, et al. (2007). "Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the IL2RA region in type 1 diabetes." Nat Genet **39**(9): 1074-1082.

Lupski, J. R., J. G. Reid, et al. (2010). "Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy." N Engl J Med **362**(13): 1181-1191.

MacArthur, D. G., S. Balasubramanian, et al. (2012). "A systematic survey of loss-of-function variants in human protein-coding genes." Science **335**(6070): 823-828.

MacArthur, D. G. and C. Tyler-Smith (2010). "Loss-of-function variants in the genomes of healthy humans." Hum Mol Genet **19**(R2): R125-130.

Madsen, B. E. and S. R. Browning (2009). "A groupwise association test for rare mutations using a weighted sum statistic." PLoS Genet **5**(2): e1000384.

Manolio, T. A., F. S. Collins, et al. (2009). "Finding the missing heritability of complex diseases." Nature **461**(7265): 747-753.

Margulies, E. H., M. Blanchette, et al. (2003). "Identification and characterization of multi-species conserved sequences." Genome Res **13**(12): 2507-2518.

Mathieson, I. and G. McVean (2012). "Differential confounding of rare and common variants in spatially structured populations." Nat Genet **44**(3): 243-246.

Mathieu-Daude, F., J. Welsh, et al. (1996). "DNA rehybridization during PCR: the 'Cot effect' and its consequences." Nucleic Acids Res **24**(11): 2080-2086.

Maxmen, A. (2011). "Exome sequencing deciphers rare diseases." Cell **144**(5): 635-637.

McSorley, H. J., S. Gaze, et al. (2011). "Suppression of inflammatory immune responses in celiac disease by experimental hookworm infection." PLoS One **6**(9): e24092.

Meresse, B., Z. Chen, et al. (2004). "Coordinated induction by IL15 of a TCR-independent NKG2D signaling pathway converts CTL into lymphokine-activated killer cells in celiac disease." Immunity **21**(3): 357-366.

Meresse, B., G. Malamut, et al. (2012). "Celiac disease: an immunological jigsaw." Immunity **36**(6): 907-919.

Meresse, B., J. Ripoche, et al. (2009). "Celiac disease: from oral tolerance to intestinal inflammation, autoimmunity and lymphomagenesis." Mucosal Immunol **2**(1): 8-23.

Metzker, M. L. (2010). "Sequencing technologies - the next generation." Nat Rev Genet **11**(1): 31-46.

Michaelson, J. J., Y. Shi, et al. (2012). "Whole-genome sequencing in autism identifies hot spots for de novo germline mutation." Cell **151**(7): 1431-1442.

Molberg, O., S. N. Mcadam, et al. (1998). "Tissue transglutaminase selectively modifies gliadin peptides that are recognized by gut-derived T cells in celiac disease (vol 4, pg 713, 1998)." Nature Medicine **4**(8): 974-974.

Momozawa, Y., M. Mni, et al. (2011). "Resequencing of positional candidates identifies low frequency IL23R coding variants protecting against inflammatory bowel disease." Nat Genet **43**(1): 43-47.

Monsuur, A. J., P. I. de Bakker, et al. (2008). "Effective detection of human leukocyte antigen risk alleles in celiac disease using tag single nucleotide polymorphisms." PLoS One **3**(5): e2270.

Morris, A. P. and E. Zeggini (2010). "An evaluation of statistical approaches to rare variant analysis in genetic association studies." Genet Epidemiol **34**(2): 188-193.

Mungall, A. J., S. A. Palmer, et al. (2003). "The DNA sequence and analysis of human chromosome 6." Nature **425**(6960): 805-811.

Murray, J. A., S. B. Moore, et al. (2007). "HLA DQ gene dosage and risk and severity of celiac disease." Clin Gastroenterol Hepatol **5**(12): 1406-1412.

Musunuru, K., D. J. Rader, et al. (2010). "From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus." Nature **466**(7307): 714-U712.

Nachman, M. W. and S. L. Crowell (2000). "Estimate of the mutation rate per nucleotide in humans." Genetics **156**(1): 297-304.

Nakamura, K., T. Oshima, et al. (2011). "Sequence-specific error profile of Illumina sequencers." Nucleic Acids Res **39**(13): e90.

Naluai, A. T., S. Nilsson, et al. (2000). "The CTLA4/CD28 gene region on chromosome 2q33 confers susceptibility to celiac disease in a way possibly distinct from that of type 1 diabetes and other chronic inflammatory disorders." Tissue Antigens **56**(4): 350-355.

Navon, O., J. H. Sul, et al. (2013). "Rare Variant Association Testing Under Low-Coverage Sequencing." Genetics.

Neale, B. M., M. A. Rivas, et al. (2011). "Testing for an unusual distribution of rare variants." PLoS Genet **7**(3): e1001322.

Nejentsev, S., J. M. Howson, et al. (2007). "Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A." Nature **450**(7171): 887-892.

Nejentsev, S., N. Walker, et al. (2009). "Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes." Science **324**(5925): 387-389.

Nekrutenko, A. and J. Taylor (2012). "Next-generation sequencing data interpretation: enhancing reproducibility and accessibility." Nat Rev Genet **13**(9): 667-672.

Nelson, M. R., D. Wegmann, et al. (2012). "An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people." Science **337**(6090): 100-104.

Ng, P. C. and S. Henikoff (2003). "SIFT: Predicting amino acid changes that affect protein function." Nucleic Acids Res **31**(13): 3812-3814.

Ng, P. C. and E. F. Kirkness (2010). "Whole genome sequencing." Methods Mol Biol **628**: 215-226.

Ng, S. B., A. W. Bigham, et al. (2010). "Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome." Nat Genet **42**(9): 790-793.

Ng, S. B., K. J. Buckingham, et al. (2010). "Exome sequencing identifies the cause of a mendelian disorder." Nat Genet **42**(1): 30-35.

Ng, S. B., E. H. Turner, et al. (2009). "Targeted capture and massively parallel sequencing of 12 human exomes." Nature **461**(7261): 272-276.

Nica, A. C., S. B. Montgomery, et al. (2010). "Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations." PLoS Genet **6**(4): e1000895.

Nicolae, D. L., E. Gamazon, et al. (2010). "Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS." PLoS Genet **6**(4): e1000888.

Nistico, L., C. Fagnani, et al. (2006). "Concordance, disease progression, and heritability of coeliac disease in Italian twins." Gut **55**(6): 803-808.

Norton, N., P. D. Robertson, et al. (2012). "Evaluating pathogenicity of rare variants from dilated cardiomyopathy in the exome era." Circ Cardiovasc Genet **5**(2): 167-174.

Nyholt, D. R. (2000). "All LODs are not created equal." Am J Hum Genet **67**(2): 282-288.

O'Roak, B. J., P. Deriziotis, et al. (2011). "Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations." Nat Genet **43**(6): 585-589.

O'Roak, B. J., L. Vives, et al. (2012). "Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders." Science **338**(6114): 1619-1622.

Ogura, Y., D. K. Bonen, et al. (2001). "A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease." Nature **411**(6837): 603-606.

Ott, J. and A. Bhat (1999). "Linkage analysis in heterogeneous and complex traits." Eur Child Adolesc Psychiatry **8 Suppl 3**: 43-46.

Park, J. H., S. Wacholder, et al. (2010). "Estimation of effect size distribution from genome-wide association studies and implications for future discoveries." Nat Genet **42**(7): 570-575.

Parrish-Novak, J., S. R. Dillon, et al. (2000). "Interleukin 21 and its receptor are involved in NK cell expansion and regulation of lymphocyte function." Nature **408**(6808): 57-63.

Pasaniuc, B., N. Rohland, et al. (2012). "Extremely low-coverage sequencing and imputation increases power for genome-wide association studies." Nat Genet **44**(6): 631-635.

Percopo, S., M. C. Babron, et al. (2003). "Saturation of the 5q31-q33 candidate region for coeliac disease." Ann Hum Genet **67**(Pt 3): 265-268.

Percy, M. J., P. W. Furlow, et al. (2008). "A gain-of-function mutation in the HIF2A gene in familial erythrocytosis." N Engl J Med **358**(2): 162-168.

Petaros, P., S. Martelossi, et al. (2002). "Prevalence of autoimmune disorders in relatives of patients with celiac disease." Dig Dis Sci **47**(7): 1427-1431.

Phillips, P. C. (2008). "Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems." Nat Rev Genet **9**(11): 855-867.

Pierce, S. B., T. Walsh, et al. (2010). "Mutations in the DBP-deficiency protein HSD17B4 cause ovarian dysgenesis, hearing loss, and ataxia of Perrault Syndrome." Am J Hum Genet **87**(2): 282-288.

Pinto, D., A. T. Pagnamenta, et al. (2010). "Functional impact of global rare copy number variation in autism spectrum disorders." Nature **466**(7304): 368-372.

Plenge, R. M., E. A. Stahl, et al. (2010). "Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci." Nature Genetics **42**(6): 508-U556.

Polanco, I. (2008). "Celiac Disease." Journal of Pediatric Gastroenterology and Nutrition **47**: S3-S6 10.1097/MPG.1090b1013e3181818df3181815.

Polychronakos, C. (2011). "Fine points in mapping autoimmunity." Nat Genet **43**(12): 1173-1174.

Pras, E., N. Raben, et al. (1995). "Mutations in the SLC3A1 transporter gene in cystinuria." Am J Hum Genet **56**(6): 1297-1303.

Preston, M.D and F Dudbridge. (2013). "Using family-based designs for detecting rare variant disease association" *in press*

Price, A. L., G. V. Kryukov, et al. (2010). "Pooled association tests for rare variants in exon-resequencing studies." Am J Hum Genet **86**(6): 832-838.

Pritchard, J. K. (2001). "Are rare variants responsible for susceptibility to complex diseases?" Am J Hum Genet **69**(1): 124-137.

Purcell, S., S. S. Cherny, et al. (2003). "Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits." Bioinformatics **19**(1): 149-150.

Purcell, S., B. Neale, et al. (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses." Am J Hum Genet **81**(3): 559-575.

Purcell, S. M., N. R. Wray, et al. (2009). "Common polygenic variation contributes to risk of schizophrenia and bipolar disorder." Nature **460**(7256): 748-752.

Ramensky, V., P. Bork, et al. (2002). "Human non-synonymous SNPs: server and survey." Nucleic Acids Res **30**(17): 3894-3900.

Ratan, A., W. Miller, et al. (2013). "Comparison of sequencing platforms for single nucleotide variant calls in a human sample." PLoS One **8**(2): e55089.

Raychaudhuri, S., O. Iartchouk, et al. (2011). "A rare penetrant mutation in CFH confers high risk of age-related macular degeneration." Nat Genet **43**(12): 1232-1236.

Regalado, E. S., D. C. Guo, et al. (2011). "Exome sequencing identifies SMAD3 mutations as a cause of familial thoracic aortic aneurysm and dissection with intracranial and other arterial aneurysms." Circ Res **109**(6): 680-686.

Reich, D. E. and E. S. Lander (2001). "On the allelic spectrum of human disease." Trends Genet **17**(9): 502-510.

Riordan, J. R., J. M. Rommens, et al. (1989). "Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA." Science **245**(4922): 1066-1073.

Risch, N. (1990). "Linkage strategies for genetically complex traits. II. The power of affected relative pairs." Am J Hum Genet **46**(2): 229-241.

Risch, N. and K. Merikangas (1996). "The future of genetic studies of complex human diseases." Science **273**(5281): 1516-1517.

Rivas, M. A., M. Beaudoin, et al. (2011). "Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease." Nat Genet **43**(11): 1066-1073.

Roach, J. C., G. Glusman, et al. (2010). "Analysis of genetic inheritance in a family quartet by whole-genome sequencing." Science **328**(5978): 636-639.

Rodriguez-Rodero, S., L. Rodrigo, et al. (2006). "MHC class I chain-related gene B promoter polymorphisms and celiac disease." Hum Immunol **67**(3): 208-214.

Romanos, J., D. Barisani, et al. (2009). "Six new coeliac disease loci replicated in an Italian population confirm association with coeliac disease." J Med Genet **46**(1): 60-63.

Romeo, S., L. A. Pennacchio, et al. (2007). "Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL." Nat Genet **39**(4): 513-516.

Rybicki, B. A. and R. C. Elston (2000). "The relationship between the sibling recurrence-risk ratio and genotype relative risk." Am J Hum Genet **66**(2): 593-604.

Sanders, S. J., M. T. Murtha, et al. (2012). "De novo mutations revealed by whole-exome sequencing are strongly associated with autism." Nature **485**(7397): 237-241.

Sandilands, A., A. Terron-Kwiatkowski, et al. (2007). "Comprehensive analysis of the gene encoding filaggrin uncovers prevalent and rare mutations in ichthyosis vulgaris and atopic eczema." Nat Genet **39**(5): 650-654.

Saxonov, S., P. Berg, et al. (2006). "A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters." Proc Natl Acad Sci U S A **103**(5): 1412-1417.

Schaub, M. A., A. P. Boyle, et al. (2012). "Linking disease associations with regulatory information in the human genome." Genome Res **22**(9): 1748-1759.

Schork, N. J., S. S. Murray, et al. (2009). "Common vs. rare allele hypotheses for complex diseases." Curr Opin Genet Dev **19**(3): 212-219.

Servin, B. and M. Stephens (2007). "Imputation-based analysis of association studies: candidate regions and quantitative traits." PLoS Genet **3**(7): e114.

Shalek, A. K., R. Satija, et al. (2013). "Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells." Nature.

Shendure, J. and H. Ji (2008). "Next-generation DNA sequencing." Nat Biotechnol **26**(10): 1135-1145.

Sheridan, B. S. and L. Lefrancois (2011). "Regional and mucosal memory T cells." Nat Immunol **12**(6): 485-491.

Shug, J., P. Chen, et al. (2013). "Analysis of single-cell transcriptomes reveals gene expression states that define cellular subpopulations." Fluidigm Corporation

Sjostrom, H., K. E. Lundin, et al. (1998). "Identification of a gliadin T-cell epitope in coeliac disease: general importance of gliadin deamidation for intestinal T-cell recognition." Scand J Immunol **48**(2): 111-115.

Skovbjerg, H., G. H. Hansen, et al. (2004). "Intestinal tissue transglutaminase in coeliac disease of children and adults: ultrastructural localization and variation in expression." Scand J Gastroenterol **39**(12): 1219-1227.

Slatkin, M. (2009). "Epigenetic inheritance and the missing heritability problem." Genetics **182**(3): 845-850.

Smith, E. M., X. Wang, et al. (2006). "Comparison of linkage disequilibrium patterns between the HapMap CEPH samples and a family-based cohort of Northern European descent." Genomics **88**(4): 407-414.

Smyth, D. J., V. Plagnol, et al. (2008). "Shared and distinct genetic variants in type 1 diabetes and celiac disease." N Engl J Med **359**(26): 2767-2777.

Sollid, L. M. (2000). "Molecular basis of celiac disease." Annu Rev Immunol **18**: 53-81.

Sollid, L. M., G. Markussen, et al. (1989). "Evidence for a primary association of celiac disease to a particular HLA-DQ alpha/beta heterodimer." J Exp Med **169**(1): 345-350.

Speliotes, E. K., C. J. Willer, et al. (2010). "Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index." Nat Genet **42**(11): 937-948.

Spielman, R. S., R. E. McGinnis, et al. (1993). "Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM)." Am J Hum Genet **52**(3): 506-516.

Stahl, E. A., D. Wegmann, et al. (2012). "Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis." Nat Genet **44**(5): 483-489.

Stankiewicz, P. and J. R. Lupski (2010). "Structural variation in the human genome and its role in disease." Annu Rev Med **61**: 437-455.

Stayoussef, M., J. Benmansour, et al. (2010). "Identification of specific tumor necrosis factor-alpha-susceptible and -protective haplotypes associated with the risk of type 1 diabetes." Eur Cytokine Netw **21**(4): 285-291.

Stefansson, H., D. Rujescu, et al. (2008). "Large recurrent microdeletions associated with schizophrenia." Nature **455**(7210): 232-236.

Stephens, J. C., J. A. Schneider, et al. (2001). "Haplotype variation and linkage disequilibrium in 313 human genes." Science **293**(5529): 489-493.

Stitziel, N. O., A. Kiezun, et al. (2011). "Computational and statistical approaches to analyzing variants identified by exome sequencing." Genome Biol **12**(9): 227.

Styrkarsdottir, U., G. Thorleifsson, et al. (2013). "Nonsense mutation in the LGR4 gene is associated with several human diseases and other traits." Nature.

Sulonen, A. M., P. Ellonen, et al. (2011). "Comparison of solution-based exome capture methods for next generation sequencing." Genome Biol **12**(9): R94.

Sung, Y. J., C. C. Gu, et al. (2012). "Genotype imputation for African Americans using data from HapMap phase II versus 1000 genomes projects." Genet Epidemiol **36**(5): 508-516.

Sung, Y. J. and E. M. Wijsman (2007). "Accounting for epistasis in linkage analysis of general pedigrees." Hum Hered **63**(2): 144-152.

Szatmari, P., A. D. Paterson, et al. (2007). "Mapping autism risk loci using genetic linkage and chromosomal rearrangements." Nat Genet **39**(3): 319-328.

Tack, G. J., W. H. Verbeek, et al. (2010). "The spectrum of celiac disease: epidemiology, clinical aspects and treatment." Nat Rev Gastroenterol Hepatol **7**(4): 204-213.

Talkowski, M. E., Z. Ordulu, et al. (2012). "Clinical diagnosis by whole-genome sequencing of a prenatal sample." N Engl J Med **367**(23): 2226-2232.

Tarpey, P. S., R. Smith, et al. (2009). "A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation." Nat Genet **41**(5): 535-543.

Teare, M. and J. H. Barrett (2005). "Genetic linkage studies." Lancet **366**(9490): 1036-1044.

Telenius, H., N. P. Carter, et al. (1992). "Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer." Genomics **13**(3): 718-725.

Tenesa, A. and C. S. Haley (2013). "The heritability of human disease: estimation, uses and abuses." Nat Rev Genet **14**(2): 139-149.

Tennessen, J. A., A. W. Bigham, et al. (2012). "Evolution and functional impact of rare coding variation from deep sequencing of human exomes." Science **337**(6090): 64-69.

Thorisson, G. A., A. V. Smith, et al. (2005). "The International HapMap Project Web site." Genome Res **15**(11): 1592-1593.

Tosi, R., D. Vismara, et al. (1983). "Evidence that celiac disease is primarily associated with a DC locus allelic specificity." Clin Immunol Immunopathol **28**(3): 395-404.

Trynka, G., K. A. Hunt, et al. (2011). "Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease." Nat Genet **43**(12): 1193-1201.

Trynka, G., C. Sandor, et al. (2013). "Chromatin marks identify critical cell types for fine mapping complex trait variants." Nat Genet **45**(2): 124-130.

Trynka, G., A. Zhernakova, et al. (2009). "Coeliac disease-associated risk variants in TNFAIP3 and REL implicate altered NF-kappaB signalling." Gut **58**(8): 1078-1083.

Tsoi, L. C., S. L. Spain, et al. (2012). "Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity." Nat Genet **44**(12): 1341-1348.

Tye-Din, J. A., R. P. Anderson, et al. (2010). "The effects of ALV003 pre-digestion of gluten on immune response and symptoms in celiac disease in vivo." Clin Immunol **134**(3): 289-295.

van Belzen, M. J., B. P. Koeleman, et al. (2004). "Defining the contribution of the HLA region to cis DQ2-positive coeliac disease patients." Genes Immun **5**(3): 215-220.

Van Belzen, M. J., J. W. Meijer, et al. (2003). "A major non-HLA locus in celiac disease maps to chromosome 19." Gastroenterology **125**(4): 1032-1041.

Van Belzen, M. J., J. W. R. Meijer, et al. (2003). "A major non-HLA locus in celiac disease maps to chromosome 19." Gastroenterology **125**(4): 1032-1041.

van Heel, D. A., L. Franke, et al. (2007). "A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21." Nat Genet **39**(7): 827-829.

Van Limbergen, J., R. K. Russell, et al. (2009). "Filaggrin loss-of-function variants are associated with atopic comorbidity in pediatric inflammatory bowel disease." Inflamm Bowel Dis **15**(10): 1492-1498.

Vidal, C., J. Borg, et al. (2009). "Variants within protectin (CD59) and CD44 genes linked to an inherited haplotype in a family with coeliac disease." Tissue Antigens **73**(3): 225-235.

Visscher, P. M., W. G. Hill, et al. (2008). "Heritability in the genomics era--concepts and misconceptions." Nat Rev Genet **9**(4): 255-266.

Visscher, P. M., S. E. Medland, et al. (2006). "Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings." PLoS Genet **2**(3): e41.

Wang, J. L., X. Yang, et al. (2010). "TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing." Brain **133**(Pt 12): 3510-3518.

Wang, K., M. Li, et al. (2010). "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data." Nucleic Acids Res **38**(16): e164.

Washizuka, S., C. Kakiuchi, et al. (2003). "Association of mitochondrial complex I subunit gene NDUFV2 at 18p11 with bipolar disorder." Am J Med Genet B Neuropsychiatr Genet **120B**(1): 72-78.

Watts, C. (2004). "Class II MHC: sweetening the peptide only diet?" Cell **117**(5): 558-559.

Wetterstrand, K. "DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)." Retrieved 15th February 2013, from www.genome.gov/sequencingcosts.

Williamson, S. H., R. Hernandez, et al. (2005). "Simultaneous inference of selection and population growth from patterns of variation in the human genome." Proc Natl Acad Sci U S A **102**(22): 7882-7887.

Wu, M. C., S. Lee, et al. (2011). "Rare-variant association testing for sequencing data with the sequence kernel association test." Am J Hum Genet **89**(1): 82-93.

Xiong, M. and S. W. Guo (1997). "Fine-scale genetic mapping based on linkage disequilibrium: theory and applications." Am J Hum Genet **60**(6): 1513-1531.

Yamaguchi, T., K. Hosomichi, et al. (2011). "Exome resequencing combined with linkage analysis identifies novel PTH1R variants in primary failure of tooth eruption in Japanese." J Bone Miner Res.

Yamanouchi, J., D. Rainbow, et al. (2007). "Interleukin-2 gene variation impairs regulatory T cell function and causes autoimmunity." Nat Genet **39**(3): 329-337.

Yampolsky, L. Y., F. A. Kondrashov, et al. (2005). "Distribution of the strength of selection against amino acid replacements in human proteins." Hum Mol Genet **14**(21): 3191-3201.

Yang, J., B. Benyamin, et al. (2010). "Common SNPs explain a large proportion of the heritability for human height." Nat Genet **42**(7): 565-569.

Yeo, G. S., I. S. Farooqi, et al. (1998). "A frameshift mutation in MC4R associated with dominantly inherited human obesity." Nat Genet **20**(2): 111-112.

You, N., G. Murillo, et al. (2012). "SNP calling using genotype model selection on high-throughput sequencing data." Bioinformatics **28**(5): 643-650.

Zeggini, E. (2011). "Next-generation association studies for complex traits." Nat Genet **43**(4): 287-288.

Zeggini, E., L. J. Scott, et al. (2008). "Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes." Nat Genet **40**(5): 638-645.

Zhang, L., X. Cui, et al. (1992). "Whole genome amplification from a single cell: implications for genetic analysis." Proc Natl Acad Sci U S A **89**(13): 5847-5851.

Zhang, L., J. W. Yan, et al. (2013). "Association of TGF-beta1 +869C/T promoter polymorphism with susceptibility to autoimmune diseases: a meta-analysis." Mol Biol Rep.

Zhao, H., M. Yan, et al. (2001). "Impaired c-Jun amino terminal kinase activity and T cell differentiation in death receptor 6-deficient mice." J Exp Med **194**(10): 1441-1448.

Zheng, J., J. Yin, et al. (2013). "Meta-analysis reveals an association of STAT4 polymorphisms with systemic autoimmune disorders and anti-dsDNA antibody." Hum Immunol.

Zhernakova, A., E. A. Stahl, et al. (2011). "Meta-Analysis of Genome-Wide Association Studies in Celiac Disease and Rheumatoid Arthritis Identifies Fourteen Non-HLA Shared Loci." PLoS Genet **7**(2): e1002004.

Zhu, Q., D. Ge, et al. (2011). "A Genome-wide Comparison of the Functional Properties of Rare and Common Genetic Variants in Humans." Am J Hum Genet **88**(4): 458-468.

Zong, C., S. Lu, et al. (2012). "Genome-wide detection of single-nucleotide and copy-number variations of a single human cell." Science **338**(6114): 1622-1626.

Zuk, O., E. Hechter, et al. (2012). "The mystery of missing heritability: Genetic interactions create phantom heritability." Proc Natl Acad Sci U S A **109**(4): 1193-1198.

Appendix I

Sample information, exome in-solution capture protocol and exome sequencing summary statistics

Appendix I - A

Samples and Pedigree Information

Coeliac individuals from affected coeliac families were selected for exome sequencing. All available family samples (affected and unaffected) were genotyped on the Illumina Immunochip Infinium array.

Table 1: Coeliac samples sequenced and coeliac samples and related controls genotyped

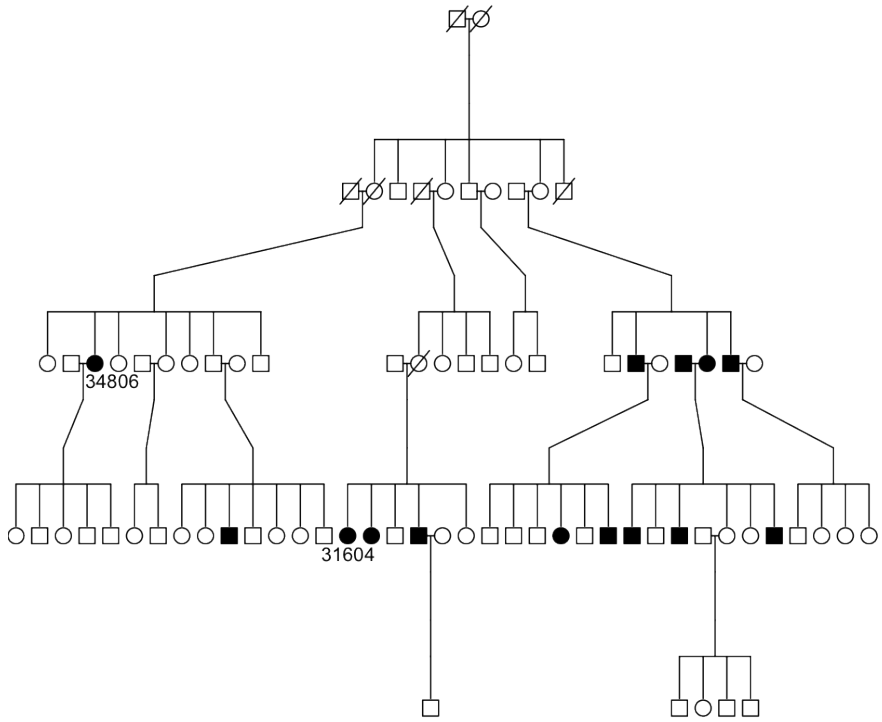
Family ID	Total number of affected individuals in family	Affected individuals sequenced	Total number genotyped (case and control)	Ethnicity
NEU4768	14/79	31604 34806	0	American
NEU4801	9/59	33165 33210 40123	0	American
NEU7017	7/49	36790 37456 38481	0	American
NEU7058	7/25	39087 39198	0	American
NEU4735	7/23	31580 35241 38794	0	American
NAL108	8/13	5846 6133	0	Swedish
DA	5	DA194 DA269	13	British
BRK	6	BRK11 BRK4	23	British
BRE	6	BRE 3	23	British
HMN	5	HMN10	14	British
BD	6	BD125	6	British
BR	4	BR88	17	British
BUT	7	BUT36	23	British
B	4	-	11	British
H	3	H74	9	British
FAM001	4	CAP152699 CAP152713 CAP200344	3	British
FAM002	4	SAL-12592-9 SAL-12706-6	2	British

SDY	13/31	SDY11 SDY20 SDY101	31	British
FAM004	3	SAL-14125-3	1	British
FAM005	6	SAL-13730-4	1	British
FAM006	4	SAL-12575-0 SAL-13281-5	2	British
FAM007	3	CAP152916 CAP200010	2	British
FAM008	8	SAL-12583-9	9	British
FAM009	5	SAL-13577-2	4	British
FAM010	7	SAL-12598-5 FAM010-4	4	British
FAM011	4	SAL-13559-1	1	British
FAM012	4	CAP152573	1	British
FAM013	3	SAL-13472-7	0	British
FAM014	6	SAL-12553-6 FAM014-6	10	British
FAM015	3	SAL-13966-5	0	British
FAM016	3	SAL-14024-1	1	British
FAM017	3	SAL-13123-0	1	British
FAM018	4	SAL-14202-9	1	British
FAM019	4	SAL-13369-2	1	British
FAM020	5	SAL-12746-0	1	British
FAM021	5	SAL-12792-1	1	British
FAM023	5	CAP152616	1	British
FAM024	7	CAP152677	1	British
FAM025	5	CAP152708	1	British
FAM026	4	CAP153119	1	British
FAM027	3	CAP152582	1	British
FAM028	5	CAP152646	1	British
FAM031	2	CAP152629	1	British
FAM033	3	CAP152730	1	British
FAM034	3	CAP152602	1	British
FAM035	3	CAP152726	1	British
FAM036	2	CAP152633	1	British
FAM037	3	CAP152825	1	British
FAM038	3	CAP153231	1	British
FAM039	4	CAP153113	1	British
FAM043	3	SAL-12544-6	1	British
FAM050	?	SAL-13357-9	1	British
FAM062	9	CUK-71848	1	British
FAM063	7	CUK-41789	7	British
FAM065	5	SAL-12847-2	1	British
SPORADIC	-	CAP152639	1	British
Total		**75**	**240**	

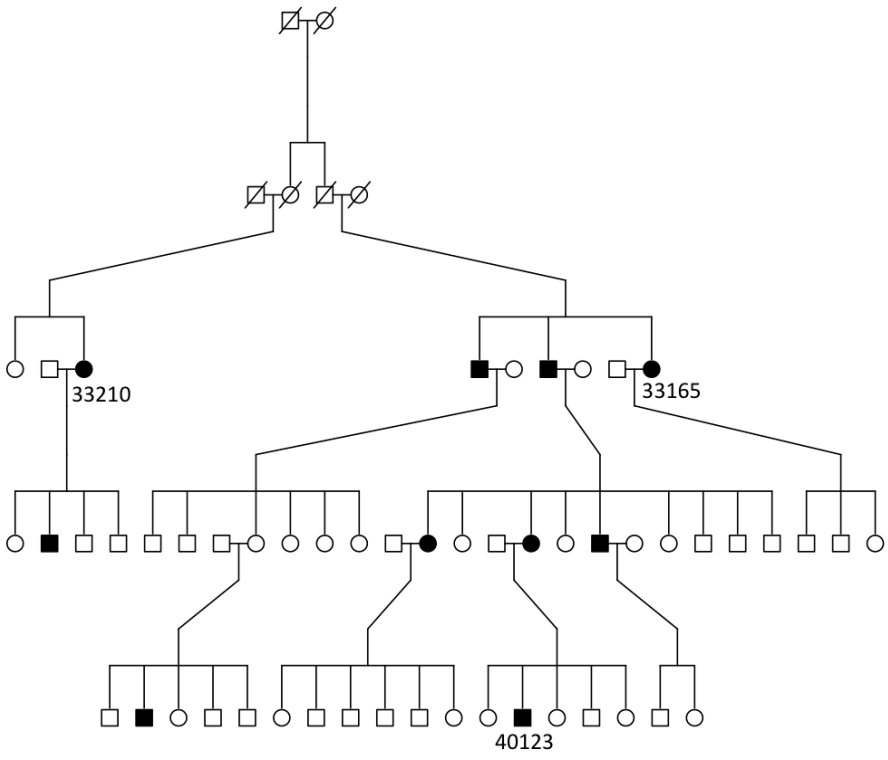
Figure 1: Pedigrees illustrating exome sequenced and genotyped samples

Samples in black are 'exome' and 'genotype'; samples in blue are 'genotype' only; pedigrees for FAM001, FAM007 and FAM010 were unavailable. HLA genotypes shown for linkage pedigrees, where 'X' denotes 'other'.

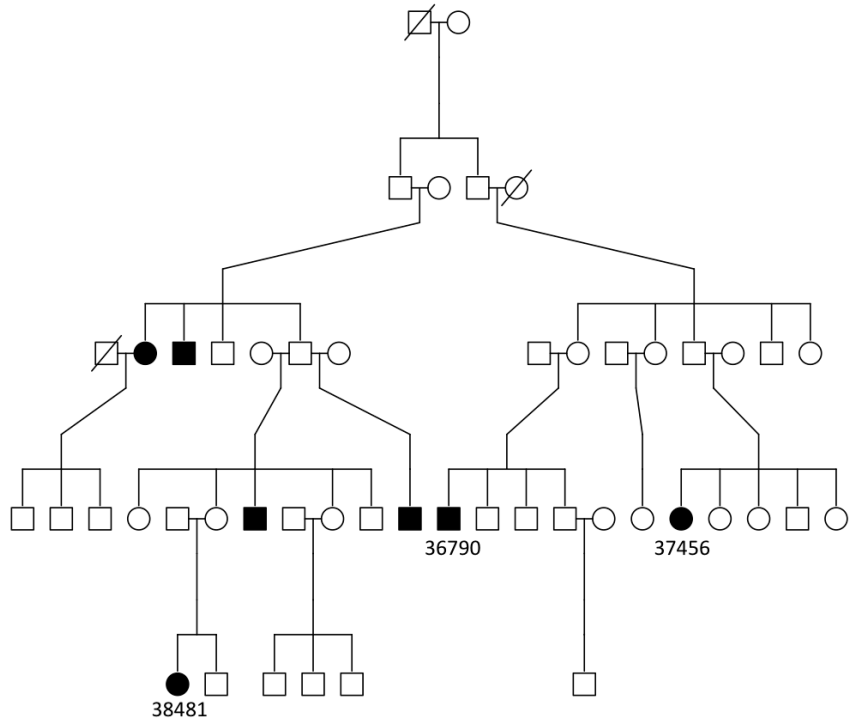
NEU4768



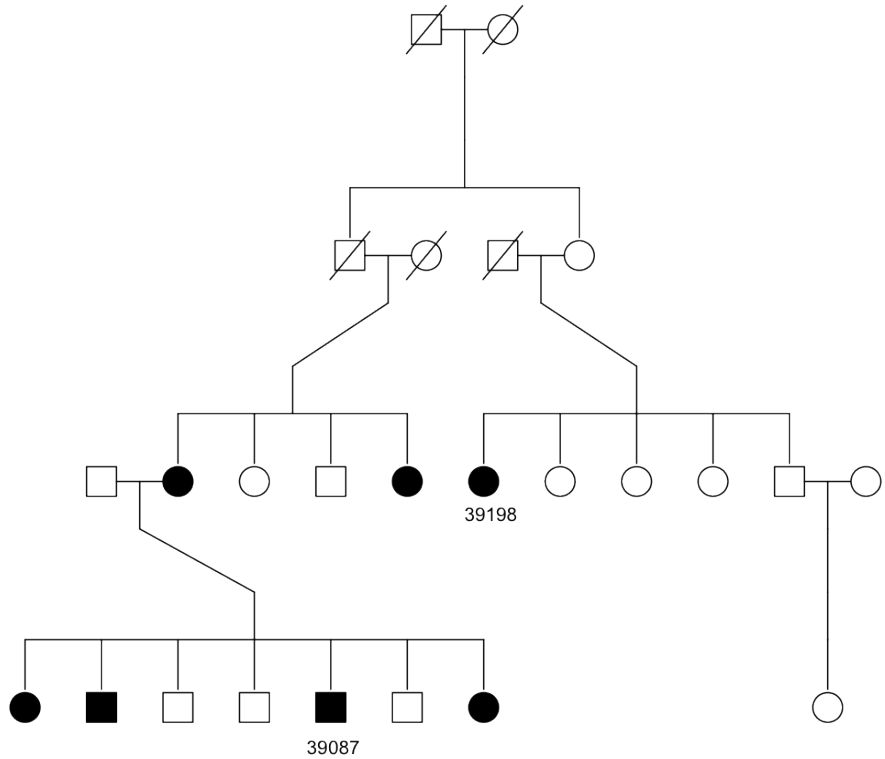
NEU4801



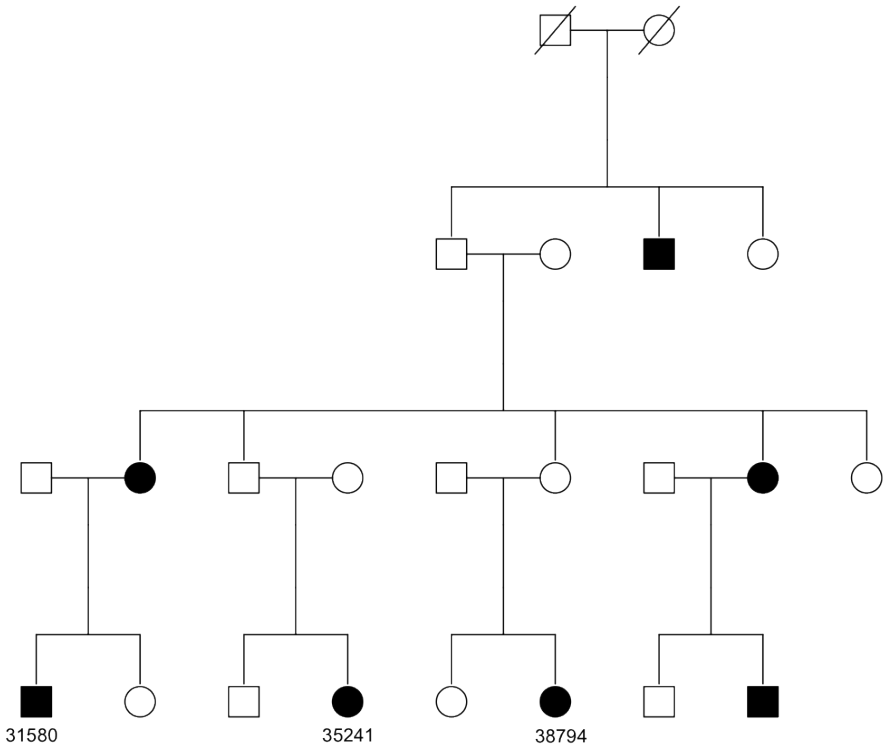
NEU7017



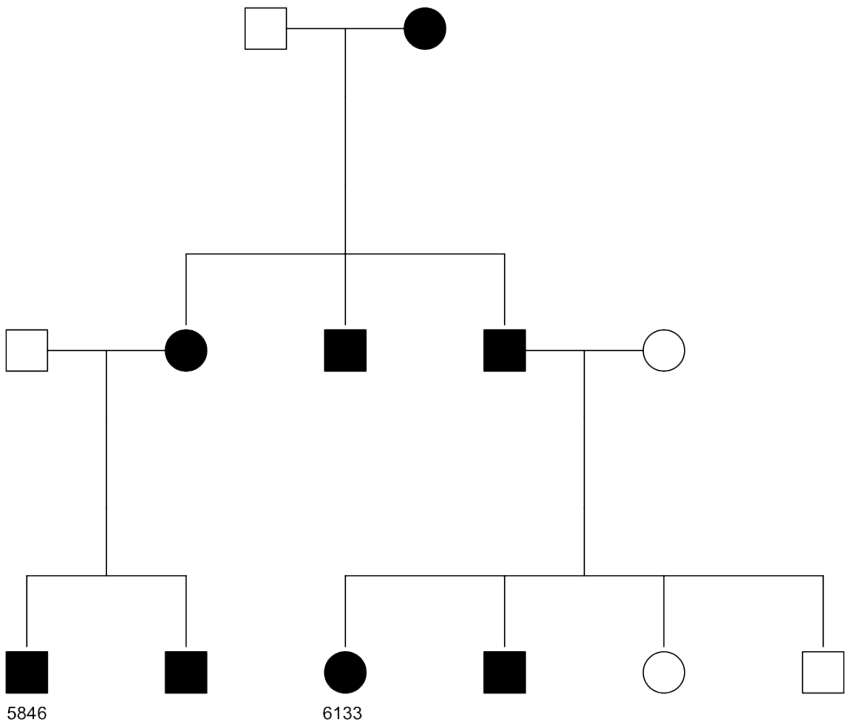
NEU7058



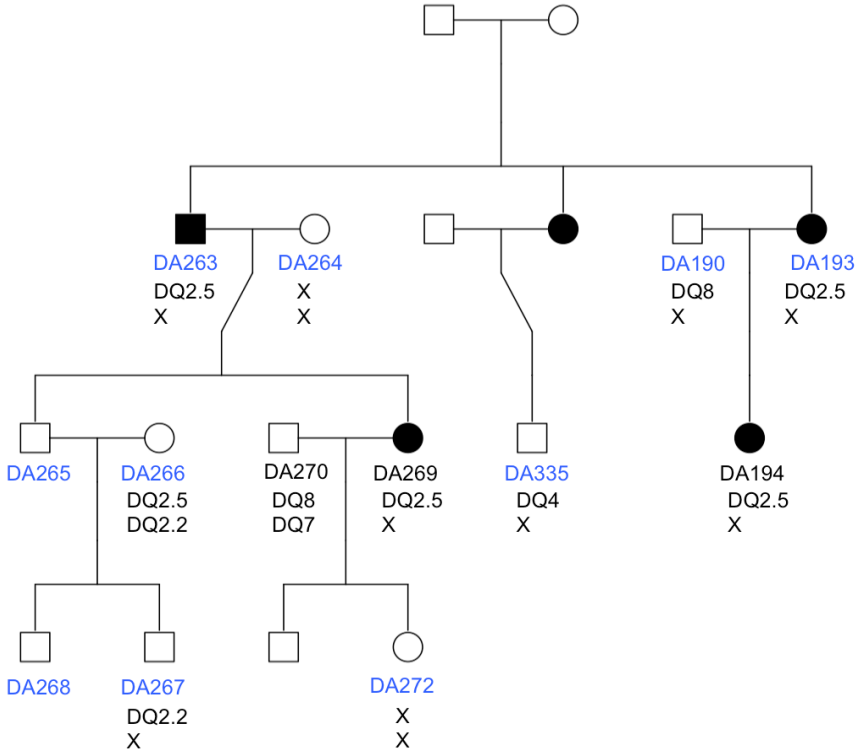
NEU4735



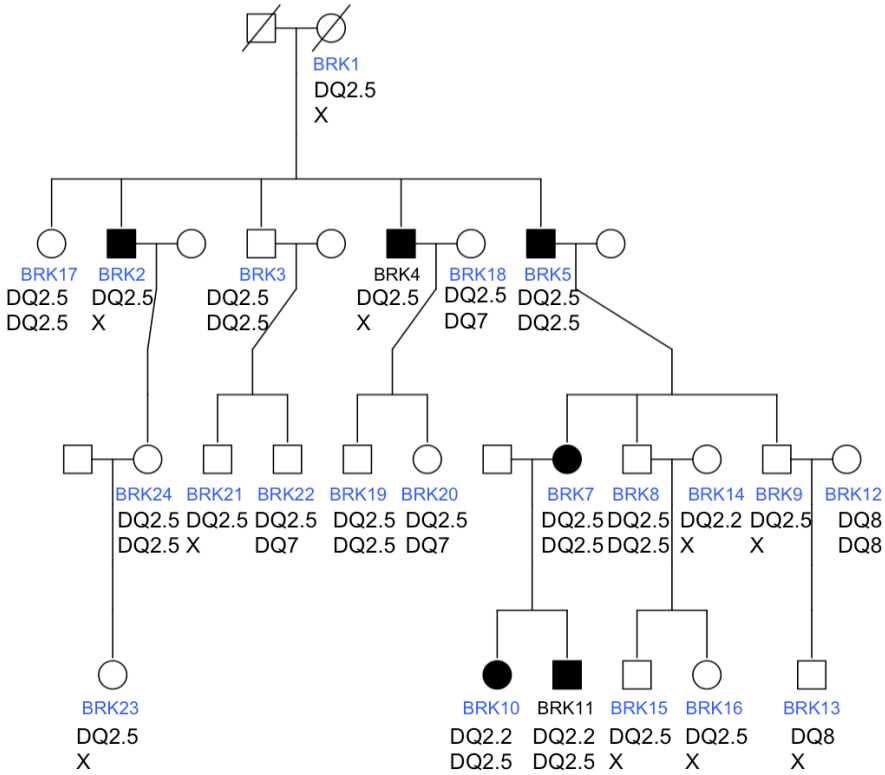
NAL108



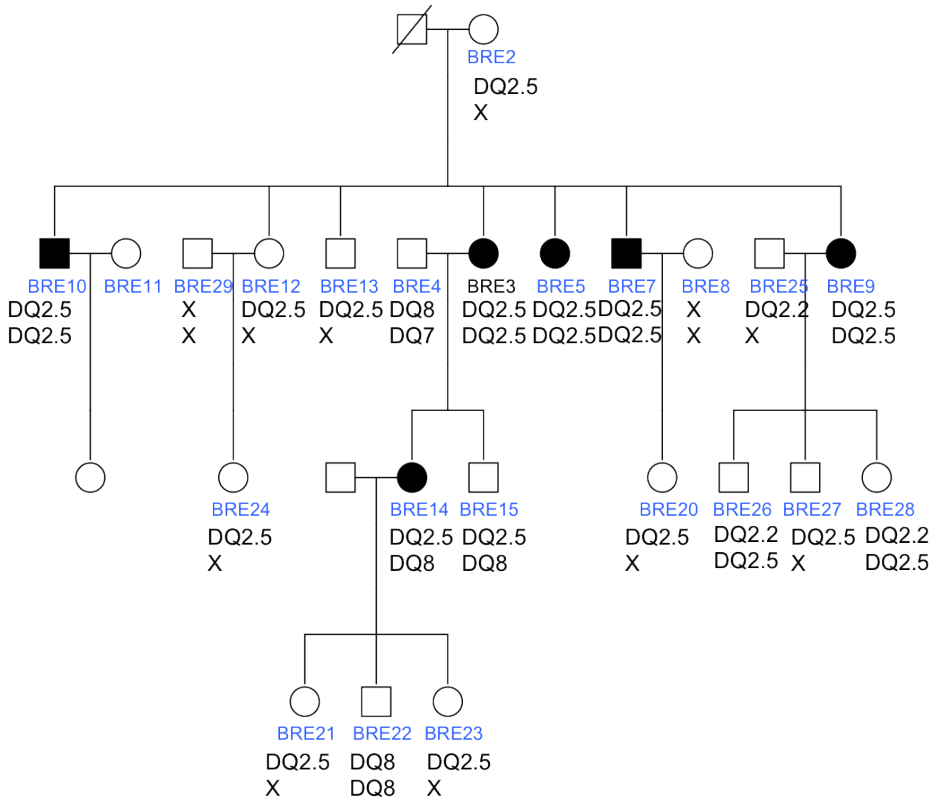
DA


DA263
DQ2.5
X
DA264
X
X
DA190
DQ8
X
DA193
DQ2.5
X
DA265
DA266
DQ2.5
DQ2.2
DA270
DQ8
DQ7
DA269
DQ2.5
X
DA335
DQ4
X
DA194
DQ2.5
X
DA268
DA267
DQ2.2
X
DA272
X
X

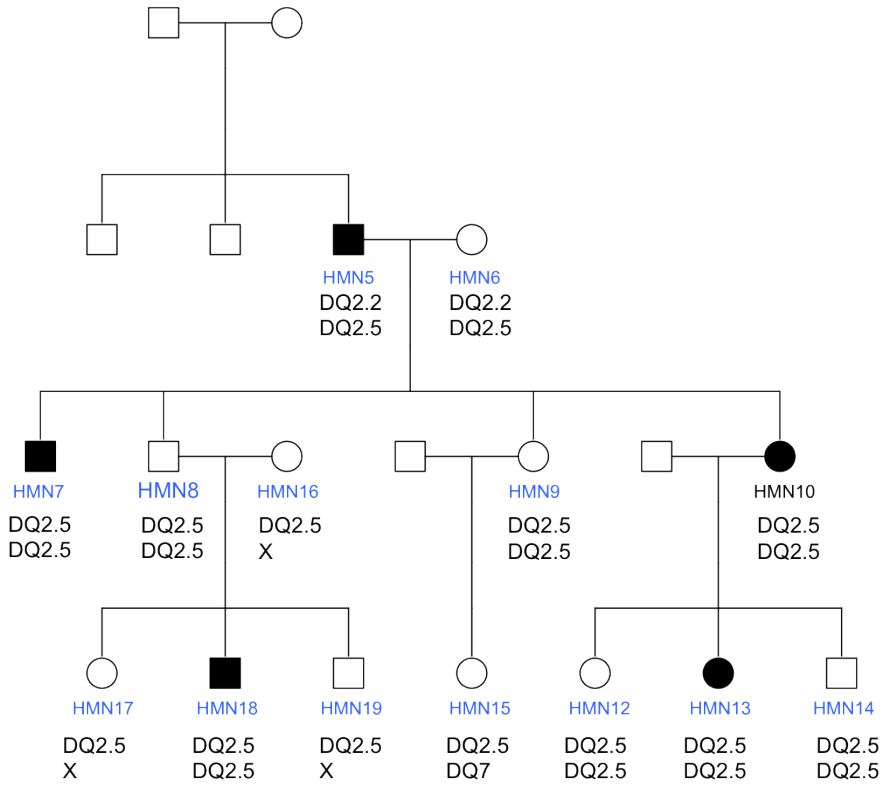
BRK


BRK1
DQ2.5
X
BRK17
DQ2.5
DQ2.5
BRK2
DQ2.5
X
BRK3
DQ2.5
DQ2.5
BRK4
DQ2.5
X
BRK18
DQ2.5
DQ7
BRK5
DQ2.5
DQ2.5
BRK24
DQ2.5
DQ2.5
BRK21
DQ2.5
X
BRK22
DQ2.5
DQ7
BRK19
DQ2.5
DQ2.5
BRK20
DQ2.5
DQ7
BRK7
DQ2.5
DQ2.5
BRK8
DQ2.5
DQ2.5
BRK14
DQ2.2
X
BRK9
DQ2.5
X
BRK12
DQ8
DQ8
BRK23
DQ2.5
X
BRK10
DQ2.2
DQ2.5
BRK11
DQ2.2
DQ2.5
BRK15
DQ2.5
X
BRK16
DQ2.5
X
BRK13
DQ8
X

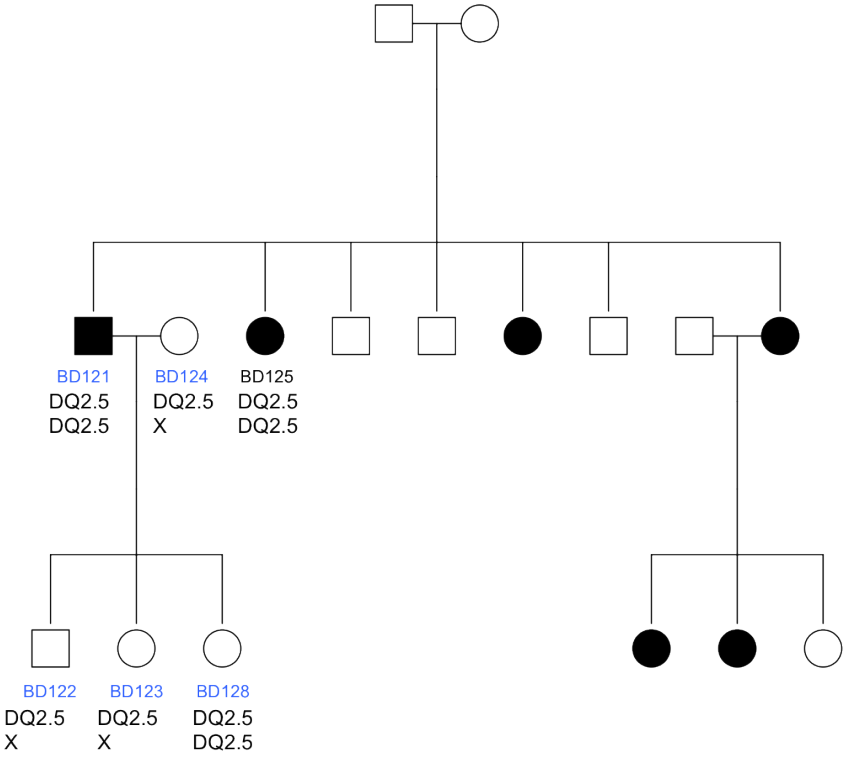
BRE


BRE2
DQ2.5
X
BRE10
DQ2.5
DQ2.5
BRE11
BRE29
X
X
BRE12
DQ2.5
X
BRE13
DQ2.5
X
BRE4
DQ8
DQ7
BRE3
DQ2.5
DQ2.5
BRE5
DQ2.5
DQ2.5
BRE7
DQ2.5
DQ2.5
BRE8
X
X
BRE25
DQ2.2
X
BRE9
DQ2.5
DQ2.5
BRE24
DQ2.5
X
BRE14
DQ2.5
DQ8
BRE15
DQ2.5
DQ8
BRE20
DQ2.5
X
BRE26
DQ2.2
DQ2.5
BRE27
DQ2.5
X
BRE28
DQ2.2
DQ2.5
BRE21
DQ2.5
X
BRE22
DQ8
DQ8
BRE23
DQ2.5
X

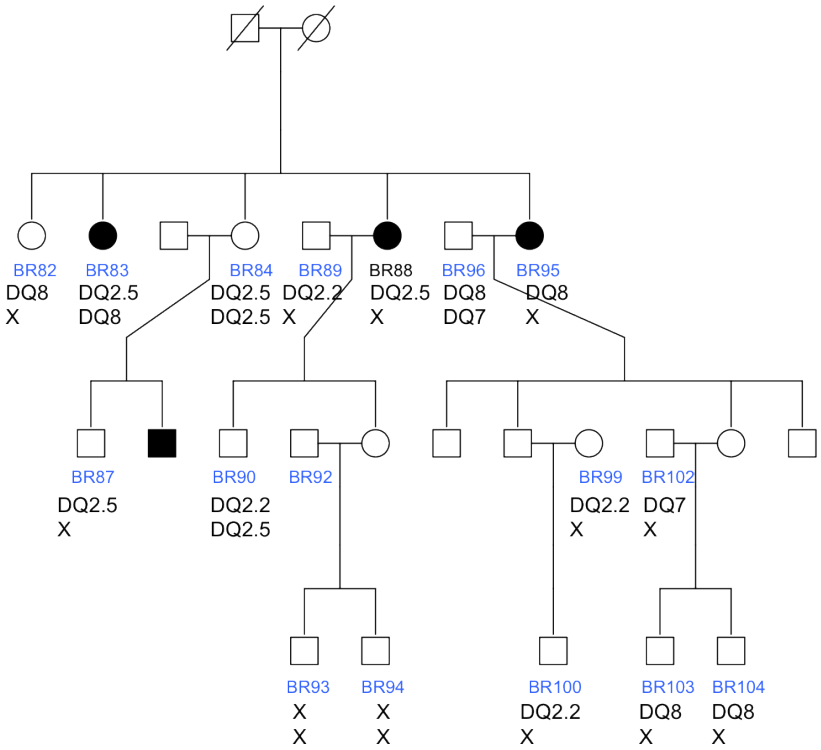
HMN


HMN5
DQ2.2
DQ2.5
HMN6
DQ2.2
DQ2.5
HMN7
DQ2.5
DQ2.5
HMN8
DQ2.5
DQ2.5
HMN16
DQ2.5
X
HMN9
DQ2.5
DQ2.5
HMN10
DQ2.5
DQ2.5
HMN17
DQ2.5
X
HMN18
DQ2.5
DQ2.5
HMN19
DQ2.5
X
HMN15
DQ2.5
DQ7
HMN12
DQ2.5
DQ2.5
HMN13
DQ2.5
DQ2.5
HMN14
DQ2.5
DQ2.5

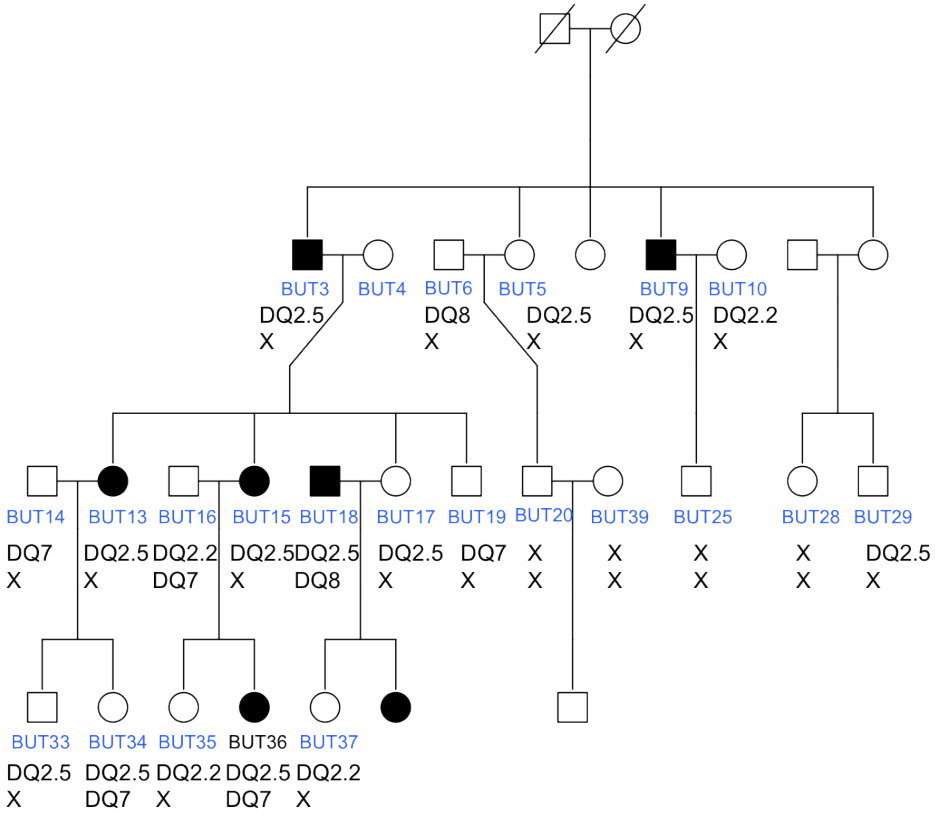
BD


BD121
DQ2.5
DQ2.5
BD124
DQ2.5
X
BD125
DQ2.5
DQ2.5
BD122
DQ2.5
X
BD123
DQ2.5
X
BD128
DQ2.5
DQ2.5

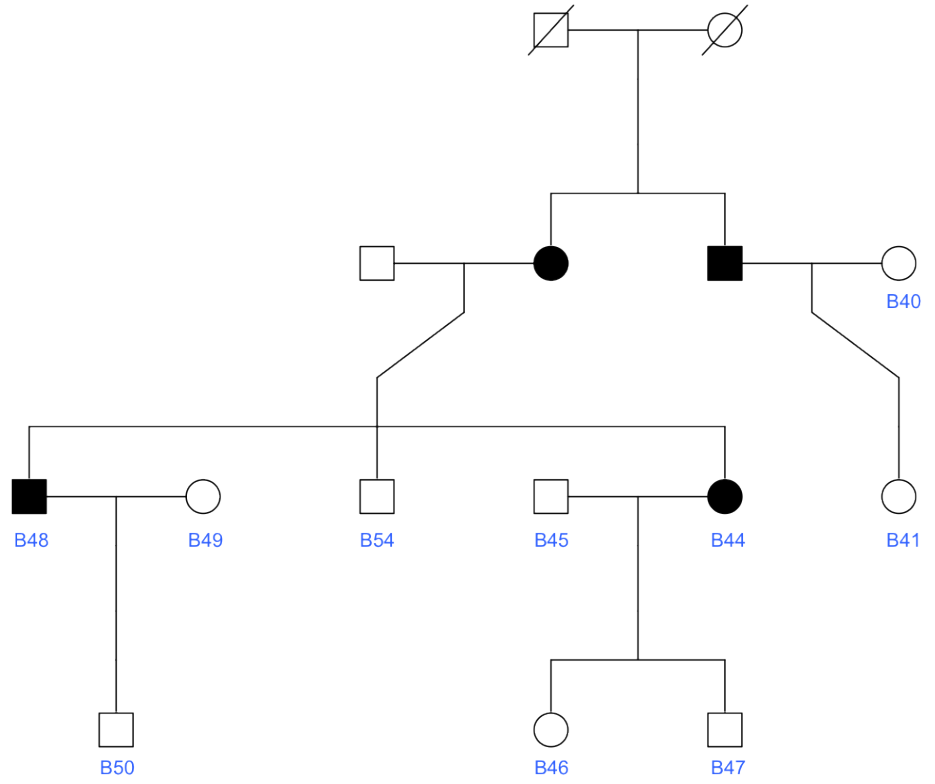
BR


BR82
DQ8
X
BR83
DQ2.5
DQ8
BR84
DQ2.5
DQ2.5
BR89
DQ2.2
X
BR88
DQ2.5
X
BR96
DQ8
DQ7
BR95
DQ8
X
BR87
DQ2.5
X
BR90
DQ2.2
DQ2.5
BR92
BR99
DQ2.2
X
BR102
DQ7
X
BR93
X
X
BR94
X
X
BR100
DQ2.2
X
BR103
DQ8
X
BR104
DQ8
X

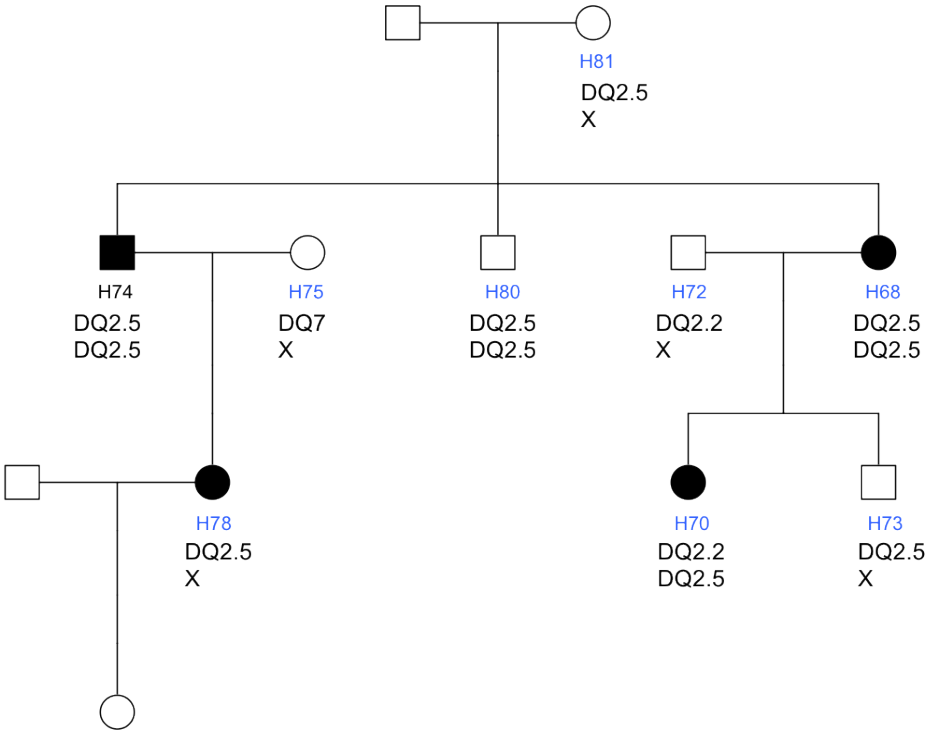
BUT


BUT3
DQ2.5
X
BUT4
BUT6
DQ8
X
BUT5
DQ2.5
X
BUT9
DQ2.5
X
BUT10
DQ2.2
X
BUT14
DQ7
X
BUT13
DQ2.5
X
BUT16
DQ2.2
DQ7
BUT15
DQ2.5
X
BUT18
DQ2.5
DQ8
BUT17
DQ2.5
X
BUT19
DQ7
X
BUT20
X
X
BUT39
X
X
BUT25
X
X
BUT28
X
X
BUT29
DQ2.5
X
BUT33
DQ2.5
X
BUT34
DQ2.5
DQ7
BUT35
DQ2.2
X
BUT36
DQ2.5
DQ7
BUT37
DQ2.2
X

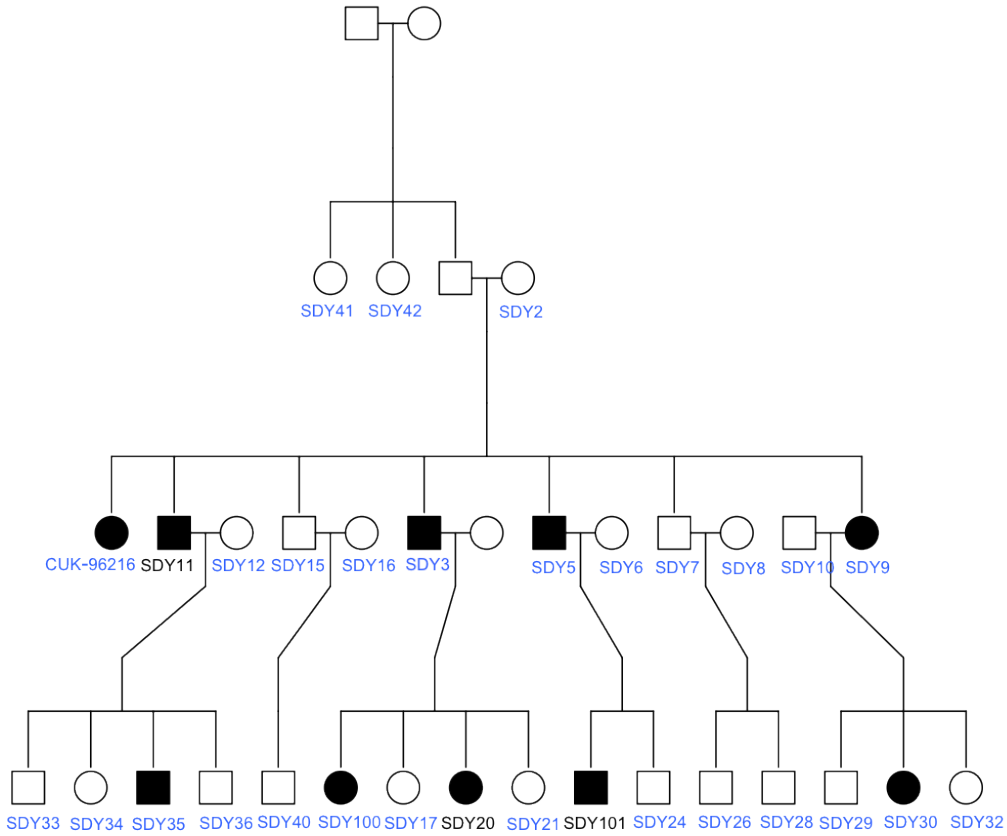
B


B40
B48
B49
B54
B45
B44
B41
B50
B46
B47

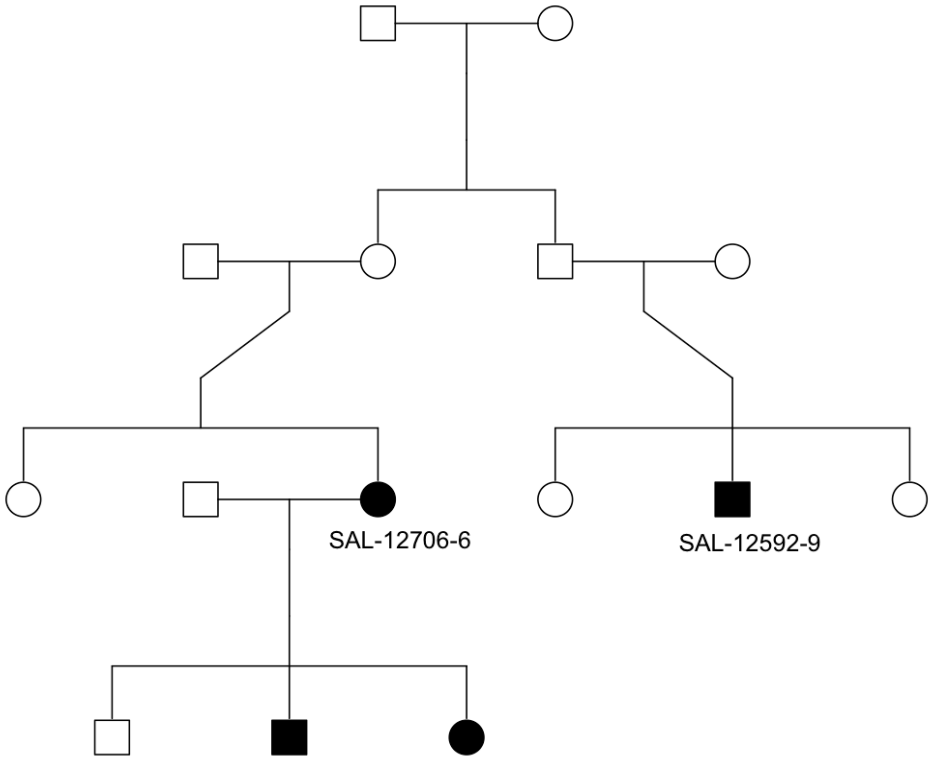
H


H81
DQ2.5
X
H74
DQ2.5
DQ2.5
H75
DQ7
X
H80
DQ2.5
DQ2.5
H72
DQ2.2
X
H68
DQ2.5
DQ2.5
H78
DQ2.5
X
H70
DQ2.2
DQ2.5
H73
DQ2.5
X

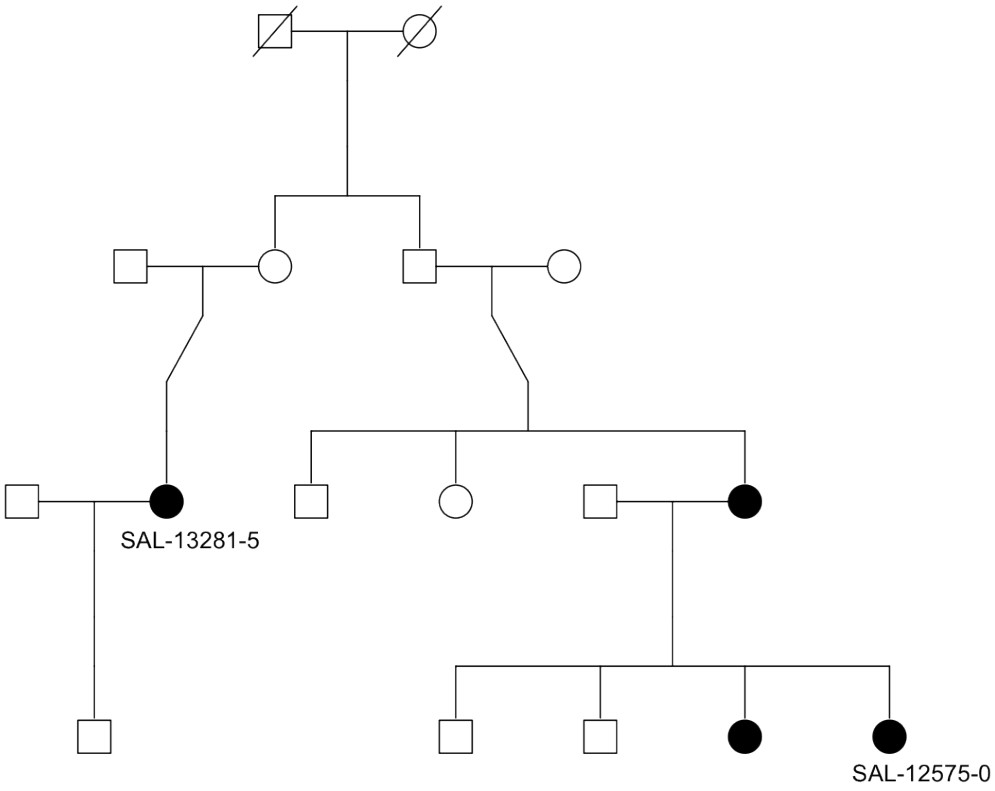
SDY


SDY41
SDY42
SDY2
CUK-96216
SDY11
SDY12
SDY15
SDY16
SDY3
SDY5
SDY6
SDY7
SDY8
SDY10
SDY9
SDY33
SDY34
SDY35
SDY36
SDY40
SDY100
SDY17
SDY20
SDY21
SDY101
SDY24
SDY26
SDY28
SDY29
SDY30
SDY32

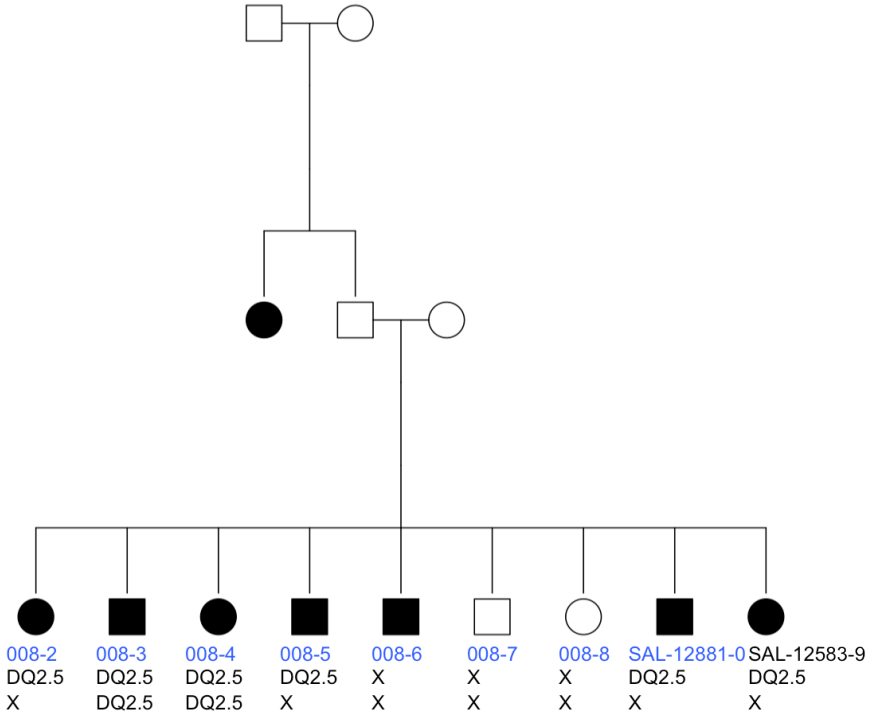
FAM002



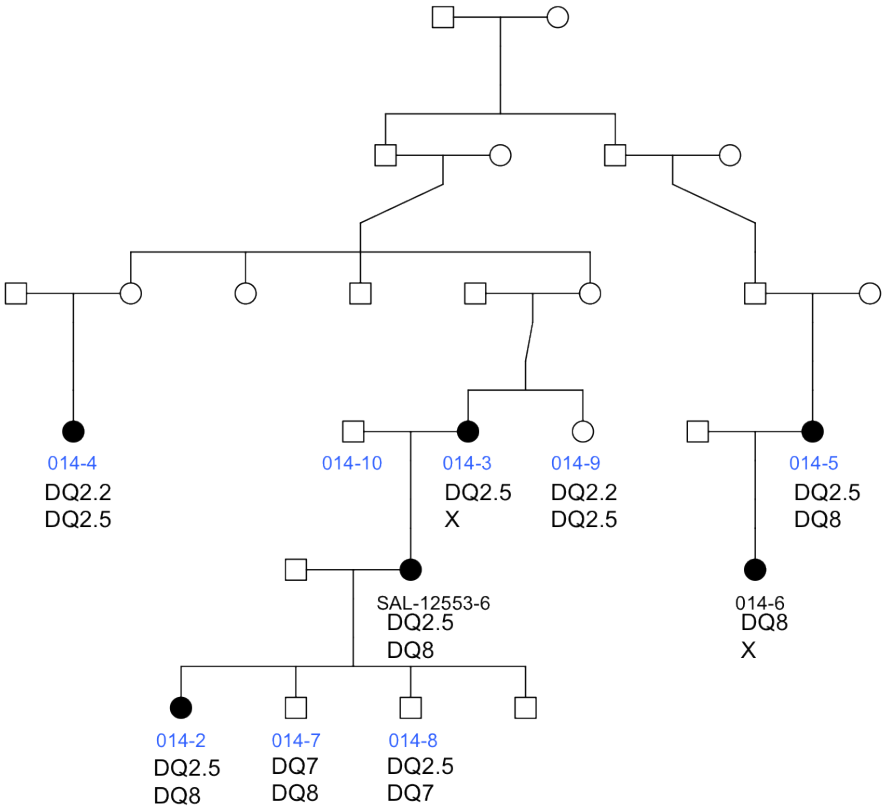
FAM006



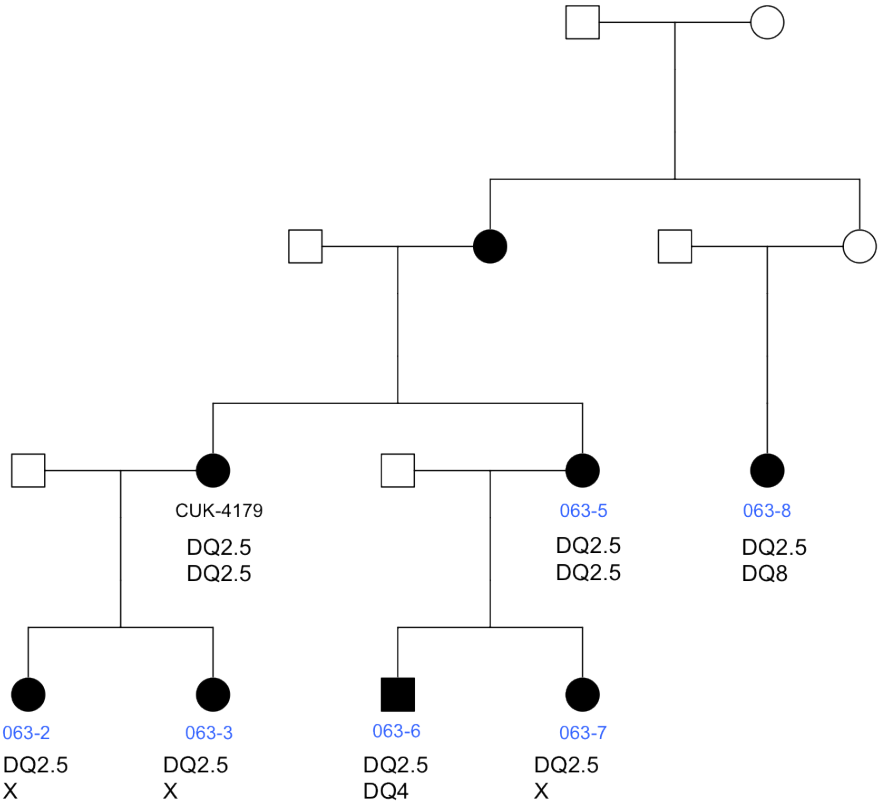
FAM008



FAM014



FAM063


CUK-4179
DQ2.5
DQ2.5
063-5
DQ2.5
DQ2.5
063-8
DQ2.5
DQ8
063-2
DQ2.5
X
063-3
DQ2.5
X
063-6
DQ2.5
DQ4
063-7
DQ2.5
X

Appendix I - B

Preparation of Illumina library prior to solution capture (EZ Exome System) with no pre hybridisation PCR

This protocol uses Sigma made paired-end PCR primers and library preparation components from New England Biosciences for exome sequence capture library preparaton and subsequent Illumina GAII paired-end sequencing. PCR is carried out post hybridisation (sequence capture) only.

Reagents/Kits

Company	Product Number	Concentration given
NEB		
Klenow Enzyme	M0210S	5U/µl, 200U per kit
Klenow Buffer (NEB buffer II)	B7002S	10x, 6ml
Klenow fragment 3-5exo	M0212L	5U/µl, 1000U per kit
T4 DNA Ligase buffer with 10mMATP	B0202S	10x (use neat), 6ml
T4 DNA Polymerase	M0203L	3U/µl, 750U
DNA ligase and buffer kit	M2200L	2000U/µl (ligase); 150 reactions
dNTP mix	N0447S	10mM, 800µl
T4 PNK	M0201L	10,000U/ml, 2500U per kit
Phusion® High-Fidelity PCR Master Mix with HF Buffer	F-531L	500 PCR reactions worth
Amersham		
dATP	28-4065-01	100mM, 25µMol pack, use at 1mM
Invitrogen		
SYBR Green I dye	S-7563	10,000X, need at 1X
Human Cot-1 DNA	15279011	1 mg/ml in 10 mM Tris-HCl (pH 7.4), 1 mM EDTA)

Dynabeads M-270 Streptavidin	653-05 653-06	2ml 10ml
Beckman Coulter		
AMPure XP beads - 5ml	A63880	5ml
Agencourt SPRIStand™	A29182	Magnetic 6-tube Stand
Qiagen		
QIAquick PCR Purification Kit	28106	250 clean ups
QIAquick MinElute PCR Purification Kit	28005	50 clean ups
Sigma		
Trizma hydrochloride	T2913-1L	1M
Sodium Chloride Solution	71386	5M
EDTA solution	E7889-100ML	0.5M

Components supplied

1. SeqCap EZ Exome Library 4 capture kit or 48 capture kit

Upon arrival, aliquot and store the SeqCap EZ Exome Library

a. If frozen, thaw the Exome Library on ice
b. Vortex the Exome library for 3 seconds
c. Centrifuge the tube of the Exome Library at 10,000 x g for 1 minute
d. Aliquot the Exome Library into 100ng single-use aliquots (4.5 µl/aliquot) in 0.2ml PCR tubes and store at -20 °C until use.

Note: *The Exome Library should not undergo multiple freeze/thaw cycles. To avoid this, it is recommended that the library is aliquoted into single-use volumes to prevent damage from successive freeze/thaw cycles*

Primers and Adapters

Order PCR primers below from Sigma Aldrich; HPLC grade and dry, and resuspend in water at given volumes for 100 µM (stock). Dilute an aliquot (working stock) of **PCR primers** to concentrations given below:

Component	Concentration	Sequence
Homemade PCR Primer PE 1.0	100 µM	5'- AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC* T -3'
Homemade PCR Primer PE 2.0	100 µM	5'-CAA GCA GAA GAC GGC ATA CGA GAT CGG TCT CGG CAT TCC TGC TGA ACC GCT CTT CCG ATC* T -3'
Homemade PE adapter 1	15 µM	5'- ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TC*T-3'
Homemade PE adapter 2	15 µM	5'-[Phos]-GAT CGG AAG AGC GGT TCA GCA GGA ATG CCG AG-3'
PE Hyb Enhancing (PE-HE) oligo 1	1000 µM	5'- AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC* T -3'
PE Hyb Enhancing (PE-HE) oligo	1000 µM	5'-CAA GCA GAA GAC GGC ATA CGA GAT CGG TCT CGG CAT TCC TGC TGA ACC GCT CTT CCG ATC* T -3'

***Phosphorothioate bond**

Adapter preparation upon arrival from Sigma:
Upon arrival from sigma Adapters must be made double stranded (i.e. annealed together) so for first time use follow instructions below, subsequent uses of the adaptors will not require this process to be carried out.

1. Mix adaptor oligonucleotides together to a final concentration of 15 µM each, in 10 mM Tris/10mM NaCl pH 7.0.

2. Anneal adapter strands in a PCR machine programmed with the following settings:
    a. Ramp at 0.5 °C/sec to 97.5 °C
    b. Hold at 97.5 °C for 150 sec
    c. 97.5 °C for 2 sec with a temperature drop of 0.1 C/cycle for 775 cycles

Note: *Store stock and working aliquots of primers and adapters at -20 °C*

<u>Fragment genomic DNA using a Covaris™</u>

Library preparation using the Illumina Paired-End Sample Prep kit requires 1-5µg of genomic DNA. Aim to fragment DNA into <1 Kb sizes.

1. Aliquot 5 µg (by picogreen concentration; see Picogreening of DNA SOP) of genomic DNA.
2. Make each 5 µg aliquot to a total volume of 80 µl with water.

3. Mix and transfer each aliquot to a new Micro tube (6mm x 16mm) AFA fibre vial.
4. Seal the tubes using a Metal crimp 8mm seal cap and crimping tube.
5. Optimise shear settings per Covaris instrument.
    a. For the following settings, size distribution obtained is between 75bp and 1000 bp and the peak size distribution is between 200-300bp: duty cycle 10%, intensity 5, cycle/bust 200, time 140s
6. Remove vials from the machine and open; on ice transfer into a fresh 1.5 ml lo-bind tube with a pipette.
7. Run 1 µl on Bioanalyzer 7500 chip for size confirmation; aim to obtain 250-300bp peaks.
8. Clean up samples with Qiagen QIAquick PCR Purification kit; elute in **30 µl** of Buffer EB into new 1.5ml lo-bind tubes.

Note: *Maximum DNA input per Qiagen clean up column is 10 µg*
Stop Point: fragmented samples can be frozen at -20 °C until library preparation

End Repair

1. Prepare end-repair reaction mix (reaction mix volume given per sample concentrations as supplied by NEB)

▪ Fragmented DNA	30 µl
▪ Molecular grade Water	46 µl
▪ T4 DNA ligase buffer	10 µl
▪ dNTP's Mix	4 µl
▪ T4 DNA Polymerase	5 µl
▪ Klenow enzyme	1 µl
▪ T4 PNK	5 µl
Total	100 µl

2. Pipette 70ul of Mastermix into each of 4 small PCR tubes; add 30ul of cleaned fragmented sample into each tube. Put cap on. Gently mix with pipette.
3. Incubate in thermal cycler for 30 minutes @ 20^{o}C
4. Follow the protocol in QIAquick PCR purification kit; elute each of the 4 samples with **32 µl** of Buffer EB into new 1.5ml lo-bind tubes.

Note: *When transferring DNA into PB buffer for first step of clean up, wash out PCR tubes with PB buffer to ensure as much as possible goes into the clean up*

Add A to the 3' end

1. Prepare end-repair reaction mix (in a 1.5ml lo-bind tube) - *before proceeding dilute dATP (5 µl dATP into 495 µl EB as not used at stock concentration) aliquot out and freeze.

- End repaired DNA 32 µl
- NEB Buffer 2 5 µl
- 1mM dATP* 10 µl
- Klenow exo 3 µl

  Total 50 µl

2. Pipette 18ul of Mastermix into each of 4 small PCR tubes; add 32ul of cleaned end repaired DNA sample into each tube. Gently mix with pipette Put cap on.
3. Incubate in thermal cycler for 30 minutes @ 37 ^{o}C
4. Follow the protocol in Qiagen MinElute PCR purification kit; elute each of the 4 samples with **10 µl** of Buffer EB into new 1.5ml lo-bind tubes.

Ligate Adaptors to DNA fragment

- **Take out AMPure™ XP beads now and allow to reach room temperature**

This procedure uses a 10:1 molar ratio of adapter to DNA insert, based on a starting quantity of 5 µg of DNA before fragmentation.

1. Prepare ligation reaction mix (reaction mix given per sample)

- A tailed DNA from 3 10 µl
- DNA ligase buffer 25 µl
- Homemade PE adapter 1 (15 µM) 5 µl
- Homemade PE adapter 2 (15 µM) 5 µl
- DNA ligase 5 µl

  Total 50 µl

2. Pipette 40ul of Mastermix into each of 4 small PCR tubes; add 10ul of cleaned ligated DNA sample into each tube. Gently mix with pipette. Put cap on.

3. Incubate for 15 minutes in thermal cycler @ 20 ^{o}C

Note: *The following SPRI bead purification needs to be carried out straight away.*

SPRI bead purification

Clean ligated sample with SPRI beads, eluting in water. This gets rid of DNA <100bp and salt/ enzymes/contaminants

Materials required:

AMPure™ XP beads *NB 6 month shelf life
Agencourt SPRIStand™ - Magnetic 6-tube Stand
Pipettes and tips
1.5ml Eppendorf Lo-Bind tubes, 3 for each sample
Vortex

Heat Block @ 37°C
Molecular biology grade water
70% Ethanol
Centrifuge

Before You Begin

- Allow beads to come to room temperature for at least 30 minutes. Reagents need to be mixed well prior to use and should appear homogeneous and consistent in colour.

- Make fresh 70 % ethanol: 7 ml 100% ethanol in 3 ml molecular biology grade water.

Purification procedure:

1. Take 90 µl of SPRI beads and add to 50 µl sample in a 1.5 ml Lo-bind tube.
2. Vortex and hold at room temperature for 5 minutes.
3. Place tube in the magnetic rack and leave for 5 minutes or until solution is clear.
4. Carefully remove the clear solution from the tubes with a pipette and discard.
5. Dispense 700µl of 70% ethanol into each tube while in the magnetic rack taking care not to disturb the magnetic beads. Aspirate and discard ethanol.
6. Repeat the ethanol wash once again (total of two washes).
7. Dry the samples on a heat block (keep the lid of the tube open) @ 37°C for 5 minutes or until the residual ethanol has evaporated. Careful not to over dry.
8. Add 50 µl of molecular biology grade water, vortex and incubate at room temperature for 2 minutes. Spin down in centrifuge.
9. Place tubes into the magnetic rack and leave for 2-3 minutes or until sample is clear.
10. Carefully remove the water and retain in a new 1.5 ml Lo-bind tube.
11. Repeat step 8 -10 once more, retaining the water in the same 1.5 ml lo-bind tube. Total volume of eluate should be **100 µl**.
12. Centrifuge the eluate at 13,000 rpm for 10 minutes.
13. Transfer the sample to a new 1.5 ml Lo-bind tube leaving behind any precipitated beads.
14. Check size of resulting fragments using 1 µl of library using Agilent DNA 7500 chip on Bioanalyzer 2100.NB adaptors on DNA will add 90~bp to length of fragmented DNA
15. Quantify sample on Nanodrop in triplicate and take average – using this concentration make 1µg aliquots. Only one 1µg aliquot will be needed for the hybridisation, others can be frozen for future hybridizations.

Stop Point: cleaned samples can be frozen at -20 °C until exome library hybridisation

Exome Library Hybridization (skip to page 15 if not performing enrichment in house)

Step 1. Prepare for hybridization

1. Turn on a heat block to 95°C and let it equilibrate to set the temperature.
2. Remove the appropriate number of 100ng probe pool aliquots from the -20 °C freezer and allow them to thaw on ice.

Step 2. Prepare hybridization cocktail

1. Add 100 µl of 1mg/ml Cot DNA and 1 µg of SPRI bead cleaned sample library to a new 1.5ml lo-bind tube.
2. Add 1 µl of each 1000 µM PE-HE1 and PE-HE2 Oligo's to the sample library plus Cot DNA.
3. Close the tube's lid and make a hole in the top of the tube's cap with a small needle
4. Dry sample in a SpeedVac on high heat (60 °C) – careful not to over-dry as DNA will not resuspend thoroughly.

Note: *Denaturation of the DNA with high heat is not problematic after linker ligation because the hybridization utilizes single-stranded DNA. This step may take 30 min or longer.*

5. To each dried-down library/COT DNA sample add:
    a. 7.5 µl 2x SC Hybridization Buffer (Nimblegen provided)
    b. 3 µl SC Hybridization Component A (Nimblegen provided)
6. Cover the hole on tube with laboratory tape. Vortex samples for 10 sec and centrifuge at maximum speed for 10 seconds.
7. Place each sample in a 95°C heat block for 10 minutes to denature the DNA.
8. Centrifuge sample at maximum speed for 10 seconds at room temperature.
9. Transfer the entire sample to the 100ng (4.5 µl aliquot) aliquot of the Exome Library in a 0.2ml PCR tube.
10. Vortex for 3 seconds and centrifuge at maximum speed for 10 seconds.
11. Incubate in a thermocycler (with heated lid turned on) at 47 °C for 64-72 hours.

Exome Library Washing and Elution of Captures Samples

Monitor water bath temperature with a high quality reliable thermometer to ensure that the water bath temperature is 47^{O}C.

Step 1. Prepare Sequence Capture Wash Buffers

1. Dilute 10X SC Wash Buffers (I, II, and III) and 2X Stringent Wash Buffer to 1X working solutions.

Component	Amount of Buffer	Amount of PCR-Grade Water	Total Final Volume
1X Stringent Wash Buffer	10ml - 2X Stringent Wash Buffer	10ml	**20ml**
1X SC Wash Buffer I	2ml – 10X SC Wash Buffer I	18ml	**20ml**
1X SC Wash Buffer II	1ml – 10X SC Wash Buffer II	9ml	**10ml**
1X SC Wash Buffer III	1ml – 10X SC Wash Buffer III	9ml	**10ml**

2. Pre-heat the following SC Wash Buffers:
   a. 20ml of Stringent Wash Buffer heated to 47°C in a water bath
   b. 5ml of SC Wash Buffer I heated to 47°C in a water bath

Step 2. Prepare Streptavidin Dynabead Binding and Wash Buffer

1. Prepare the Streptavidin Dynabead Binding and Wash Buffer in either a 15ml or 50ml conical tube, depending on the number of captures:

Component	4 Captures*	48 Captures
1M Trizma hydrochloride	25 µl	245 µl
0.5M EDTA	5 µl	49 µl
5M NaCl	1,000 µl	9,800 µl
PCR-grade water	1,470 µl	14.406 µl
Total	**2.5ml**	**25.5ml**

*Volume adjusted for pipetting variance.

2. Vortex for 20 seconds and label the tube appropriately.
3. Store the Streptavidin Dynabead Binding and Wash Buffer at room temperature. **Buffer can be stored for up to 2 months.**

Step 3. Prepare the Streptavidin Dynabeads

1. Allow the Streptavidin Dynabeads to warm to room temperature for 30 minutes prior to use.
2. Mix the beads thoroughly by vortexing for 1 minute.
3. Aliquot 100µl of beads for each capture into a single 1.5ml tube (i.e. for 1 capture use 100µl beads and for 4 captures use 400µl beads, etc.). Enough beads for 6 captures can be prepared in a single tube.

4. Place the tube on a magnet suitable to hold 1.5ml tubes. When the liquid becomes clear (approximately 5 minutes), remove and discard the liquid being careful to leave all of the beads in the tube.
5. Add twice the initial volume of beads of Streptavidin Dynabead Binding and Wash Buffer to each tube (i.e. for 1 capture use 200µl of buffer and for 4 captures use 800µl buffer, etc.).
6. Remove the tube from the magnet and vortex for 10 seconds.
7. Place the tube back on the magnet to bind the beads. Once clear, remove and discard the liquid.
8. Repeat Steps 5 - 7 (for a total of 2 washes).
9. After removing the buffer following the second wash, resuspend the beads in 1x the original volume using the Streptavidin Dynabead Binding and Wash Buffer (i.e. for 1 capture use 100µl buffer and for 4 captures use 400µl buffer, etc.).
10. Aliquot 100µl of resuspended beads into new 0.2ml tubes.
11. Use the magnet to bind the beads. Remove and discard the liquid when clear.
12. The Streptavidin Dynabeads are now ready to bind the captured DNA. Proceed directly to the next step.

Step 4. Bind DNA to the Streptavidin Dynabeads

1. Transfer the hybridization samples to the Streptavidin Dynabeads prepared in Step 3.
2. Mix thoroughly by pipetting up and down 10 times.
3. Bind the captured sample to the beads by placing the tubes containing the beads and DNA in a thermocycler set to 47°C for 45 minutes.
4. Mix the samples by vortexing for 3 seconds at 15 minute intervals to ensure that the beads remain in suspension.

Step 5. Wash the Streptavidin Dynabeads Plus Bound DNA

1. After the 45 minute incubation, transfer the entire mixture to a 1.5ml tube.
2. Use the magnet to bind the beads. Remove and discard the liquid once clear.
3. Add 100µl of SC Wash Buffer I heated to 47°C.
4. Mix by vortexing for 10 seconds.
5. Place the tubes on the magnet to bind the beads. Remove and discard the liquid once clear.
6. Remove the tubes from the magnet and add 200µl of Stringent Wash Buffer. Pipette up and down 10 times to mix.
7. Incubate at 47°C for 5 minutes.

8. Repeat Steps 5 - 7 for a total of 2 washes with Stringent Wash Buffer.
9. Place the tubes on the magnet to bind the beads. Remove and discard the liquid once clear.
10. Add 200µl of room temperature SC Wash Buffer I and mix by vortexing for 2 minutes.
11. Place the tubes on the magnet to bind the beads. Remove and discard the liquid once clear.
12. Add 200µl of room temperature SC Wash Buffer II and mix by vortexing for 1 minute.
13. Place the tubes on the magnet to bind the beads. Remove and discard the liquid once clear.
14. Add 200µl of room temperature SC Wash Buffer III and mix by vortexing for 30 seconds.
15. Place the tubes on the magnet to bind the beads. Remove and discard the liquid once clear.
16. Remove the tubes from the magnet and add 50µl PCR-grade water to each tube of bead-bound captured sample.

Stop Point: Store the beads plus captured samples at -20 C until ready to proceed to Post Capture LM-PCR

Post capture LMPCR

1. Prepare post capture LMPCR reaction mix, 12 reactions per eluate, 2 reactions with SYBR green I dye, one negative control and one positive control, both with SYBR green I dye. Prepare reactions in 1.5 ml lo-bind tubes.

2. SYBR green I given at 10 000 X concentration; dilute 1 µl in 799 µl DMSO (1:800)

▪ Eluate	4 µl
▪ 2x Phusion Master Mix	25 µl
▪ PE primer 1.0 (25 µM)	1 µl
▪ PE primer 2.0 (25 µM)	1 µl
▪ SYBR green (diluted in DMSO 1:800)	1 µl
▪ Molecular Grade water	18µl
Total	50 µl

3. Aliquot 46µl of Mastermix in 10 lo-bind tubes (without SYBR green) and 45µl of Mastermix in 2 lo-bind tubes (with SYBR green) plus the control tubes. Add 4µl of eluate per tube. Add 1 µl of SYBR green in each of the 2 lo-bind tubes and control tubes.

4. Run PCR on the Corbett Rotor-gene RT-PCR machine, using following thermal cycling conditions:

1. 98°C 30s
2. **98°C 10s**
3. **65°C 30s**
4. **72°C 30s – collect result at this step**

Terminate reaction just before amplification reaches plateau curve.

5. Make two pools of 5 PCR reactions (those without SYBR green I); add 1,250 µl Qiagen PBI buffer to each of the two pools. Follow the protocol in QIAquick PCR purification kit, elute with 50 µl EB. Mix the two resulting pools together, so **100 µl** in total from 10 PCR reactions.

Determine the post capture LMPCR concentration and size

1. Measure concentration of library by qPCR.

2. Run Agilent Bioanalyzer DNA7500 chip

Appendix I-C

Summary Statistics for 75 coeliac exomes

Sample	Mean coverage	Number of SNPs	Total reads	Total unique reads	% unique reads	% reads on target
BD125	49.9	15314	21806499	17522198	80.4	85.8
BRK11	38.7	14792	19488200	13920438	71.4	83.3
BRK4	38.0	14319	19383798	13464140	69.5	83.2
BUT36	74.3	15748	35961107	27931526	77.7	84.6
CAP152573	54.3	15105	22269704	20809469	93.4	88.5
CAP152582	51.4	15352	21261310	19875402	93.5	87.2
CAP152602	58.2	15210	25508789	23768557	93.2	85.0
CAP152616	65.7	15019	27373757	24610549	89.9	88.4
CAP152629	45.9	14779	19367470	18213413	94	87.1
CAP152633	55.9	14913	24654132	22656262	91.9	85.4
CAP152639	45.1	12791	20487543	19492259	95.1	86.2
CAP152646	60.6	15061	26787695	25029622	93.4	85.1
CAP152658	48.4	16663	20644436	19199835	93	84.5
CAP152677	107.1	15326	52997717	42201911	79.6	87.5
CAP152699	67.3	15519	28230050	24629384	87.2	89.4
CAP152708	71.4	14840	29951044	28034390	93.6	88.2
CAP152713	112.7	15290	53114021	43554049	82	89.3
CAP152726	54.9	14888	25088339	23566243	93.9	82.4
CAP152730	40.8	15044	16227946	14934907	92	86.6
CAP152825	61.1	15345	26223092	24003151	91.5	86.9
CAP152916	69.5	15235	30987209	28714566	92.7	86.6

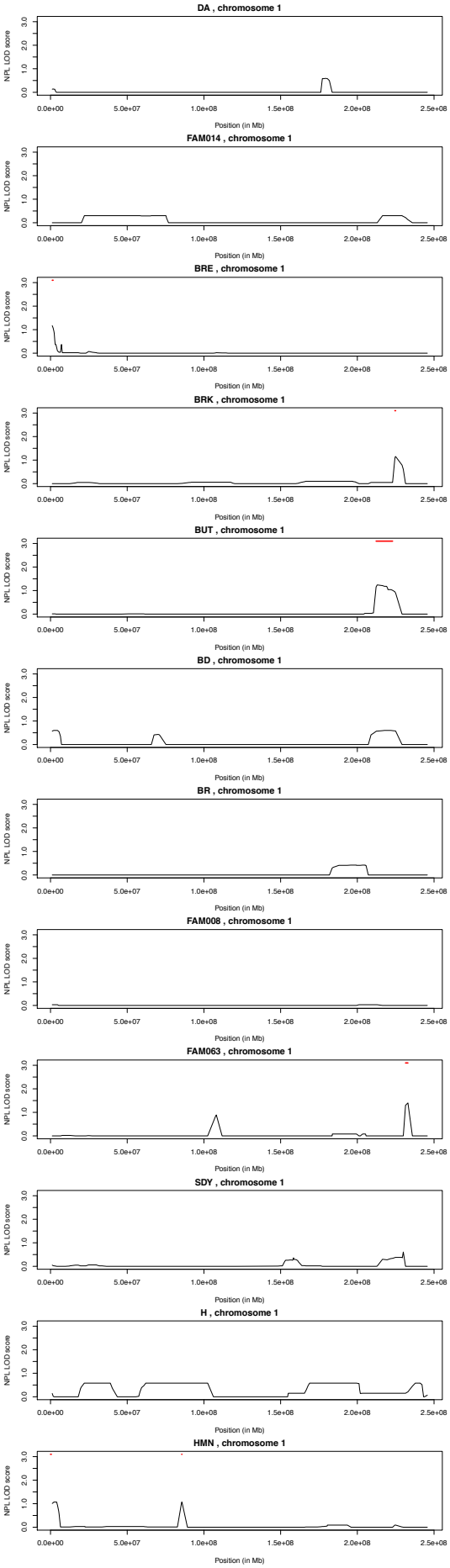
CAP153113	67.3	15152	30396359	28247009	92.9	85.0
CAP153119	65.0	15192	29112035	25688348	88.2	87.0
CAP153231	59.8	15241	26061548	24223058	92.9	85.5
CAP200010	60.2	15371	26001856	23893175	91.9	86.6
CAP200344	60.5	15614	26383621	23538963	89.2	87.5
CUK-41789	63.1	15259	34913754	28483442	81.6	75.0
CUK-71848	69.2	15620	35028306	30299232	86.5	77.8
DA194	44.8	13676	22960721	15888421	69.2	84.5
DA269	51.3	14313	22526005	17645066	78.3	86.3
FAM010-4	48.7	16638	48765253	35895980	73.6	47.3
FAM014-6	40.2	16361	43749347	29222387	66.8	47.6
H74	16.9	16505	9466873	7583305	80.1	79.7
HMN10	13.7	14263	8611430	6271787	72.8	68.7
Naluai	63.4	15612	25999072	23174204	89.1	88.2
Naluai	59.1	15493	24549860	22278612	90.7	87.2
Neuhausen4735	67.6	15228	26056990	24589360	94.4	86.3
Neuhausen4735	34.3	14893	19743950	18813543	95.3	59.9
Neuhausen4735	42.8	16164	35145165	29530782	84	51.2
Neuhausen4768	42.8	15180	25898998	24320220	93.9	59.5
Neuhausen4768	53.0	15280	20576500	19278197	93.7	86.2
Neuhausen4801	69.8	15240	28163868	25607734	90.9	88.3
Neuhausen4801	58.3	15653	24132862	21050052	87.2	85.9
Neuhausen4801	64.9	15216	26144218	23136401	88.5	87.7
Neuhausen7017	48.4	15425	21832717	19595612	89.8	78.4
Neuhausen7017	46.9	15214	17929953	15887718	88.6	87.4
Neuhausen7017	61.8	15408	24234228	21989897	90.7	87.2
Neuhausen7058	51.1	15105	19669432	18508536	94.1	85.4
Neuhausen7058	59.6	15201	22335030	20901435	93.6	88.8

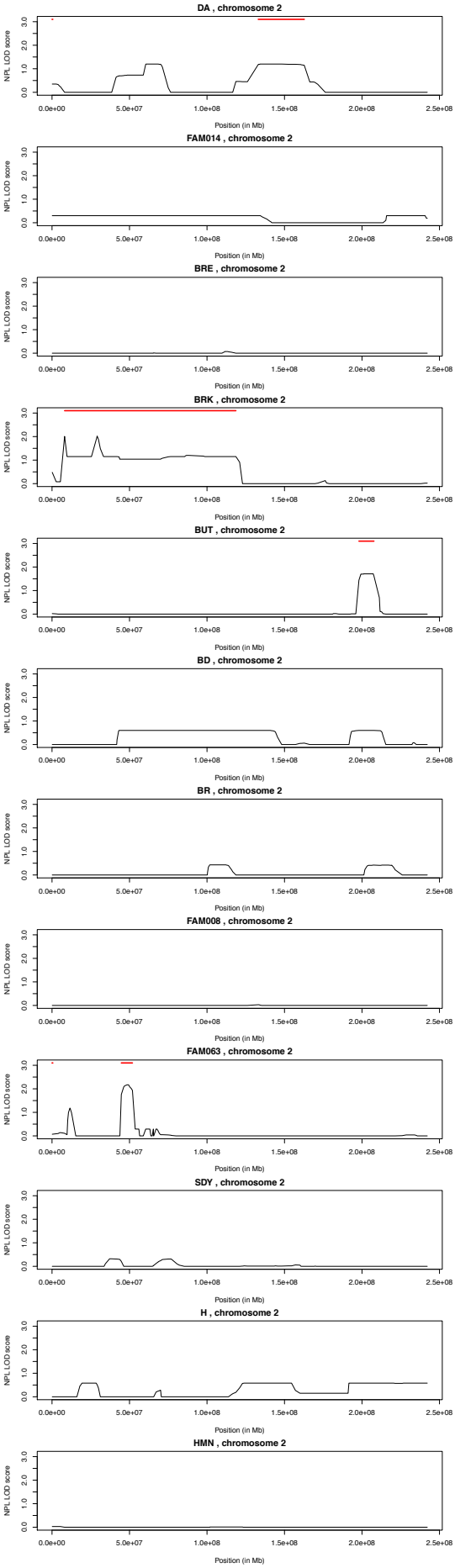
SAL-12544-6	66.4	15250	26369954	25009023	94.8	87.0
SAL-12553-6	92.0	15160	41779561	35245721	84.4	90.6
SAL-12575-0	22.2	12770	59120951	15719553	26.6	51.5
SAL-12583-9	72.9	14038	37348622	31995783	85.7	85.6
SAL-12592-9	41.8	12884	19913810	17546341	88.1	85.2
SAL-12598-5	59.7	14655	25179001	23652659	93.9	84.6
SAL-12706-6	58.6	14921	30653765	28810630	94	73.3
SAL-12746-0	40.4	12460	19672961	17778083	90.4	80.2
SAL-12792-1	66.5	14611	27731440	25907859	93.4	83.9
SAL-12847-2	52.6	16087	30728026	28824424	93.8	63.3
SAL-13093-6	102.7	14018	51141008	38837379	75.9	89.8
SAL-13123-0	37.9	12103	17947394	16605739	92.5	85.4
SAL-13281-5	84.0	14795	51658474	42532697	82.3	75.0
SAL-13357-9	28.8	14310	11087549	10610810	95.7	87.6
SAL-13359-1	36.4	12811	15527983	14628854	94.2	90.3
SAL-13369-2	35.0	13038	17239755	14750871	85.6	85.1
SAL-13472-7	33.1	12302	14658611	13469866	91.9	86.8
SAL-13477-2	63.6	15078	54726041	41168886	75.2	61.4
SAL-13730-4	38.3	13337	18527186	17142783	92.5	77.9
SAL-13966-5	34.9	12223	14623026	13531857	92.5	87.5
SAL-14024-1	33.8	12631	14677702	13580414	92.5	87.2
SAL-14125-3	38.7	13100	16706684	15391261	92.1	91.1
SAL-14202-9	40.2	13101	19139632	16817288	87.9	83.4
SDY101	47.7	15437	19426782	17179335	88.4	89.4
SDY20	60.9	14931	24104415	21420247	88.9	87.9

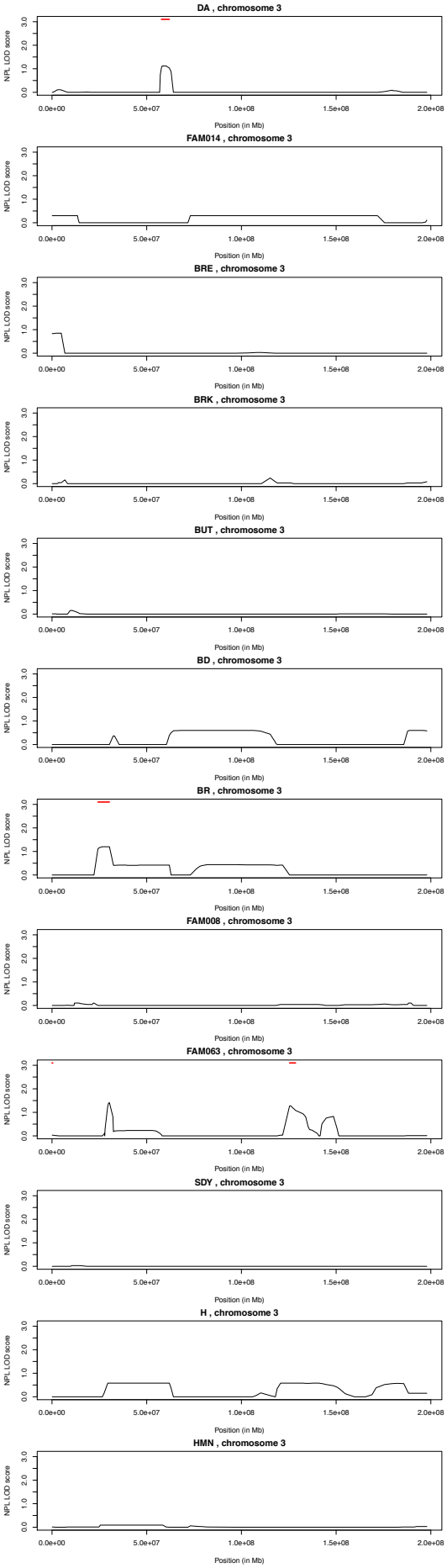
Samples in grey highlight were sequenced twice due to initial poor capture and/or sequencing run

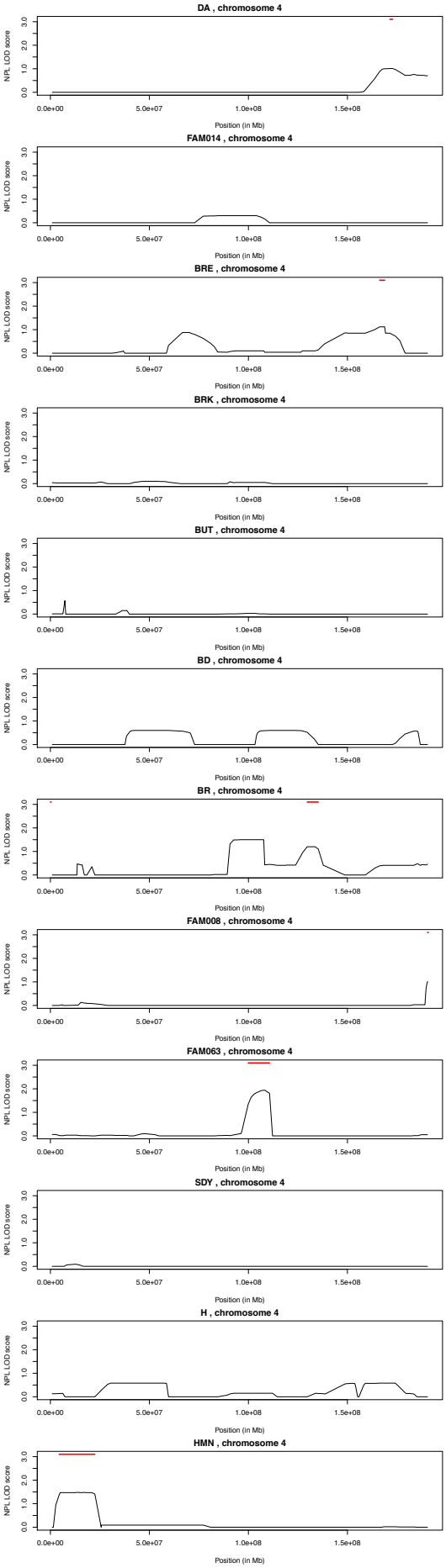
Appendix II

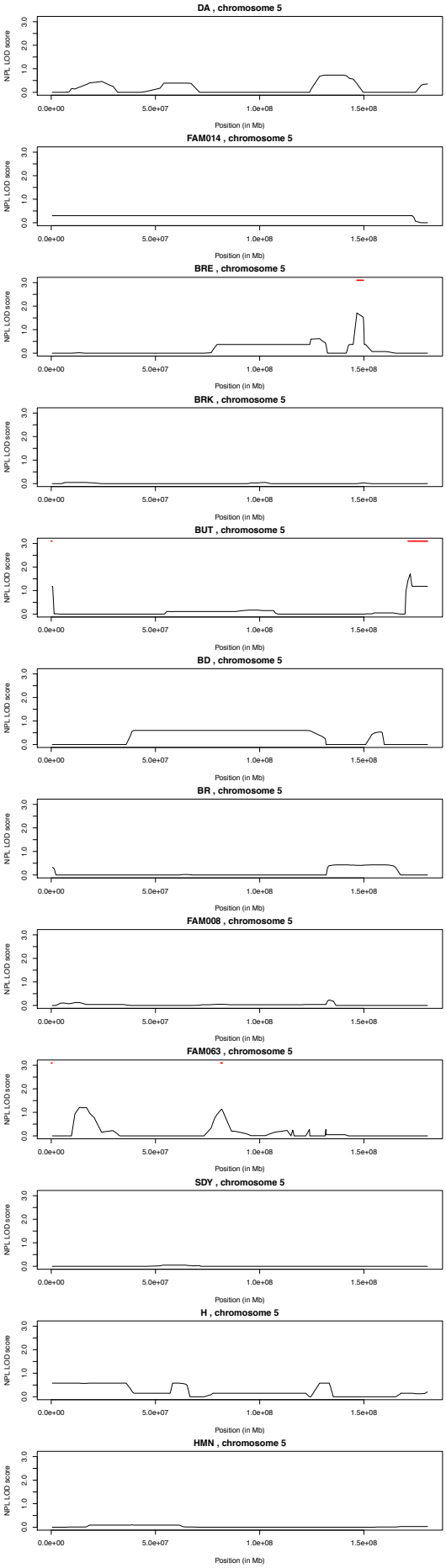
Linkage analysis graphs for 12 coeliac pedigrees

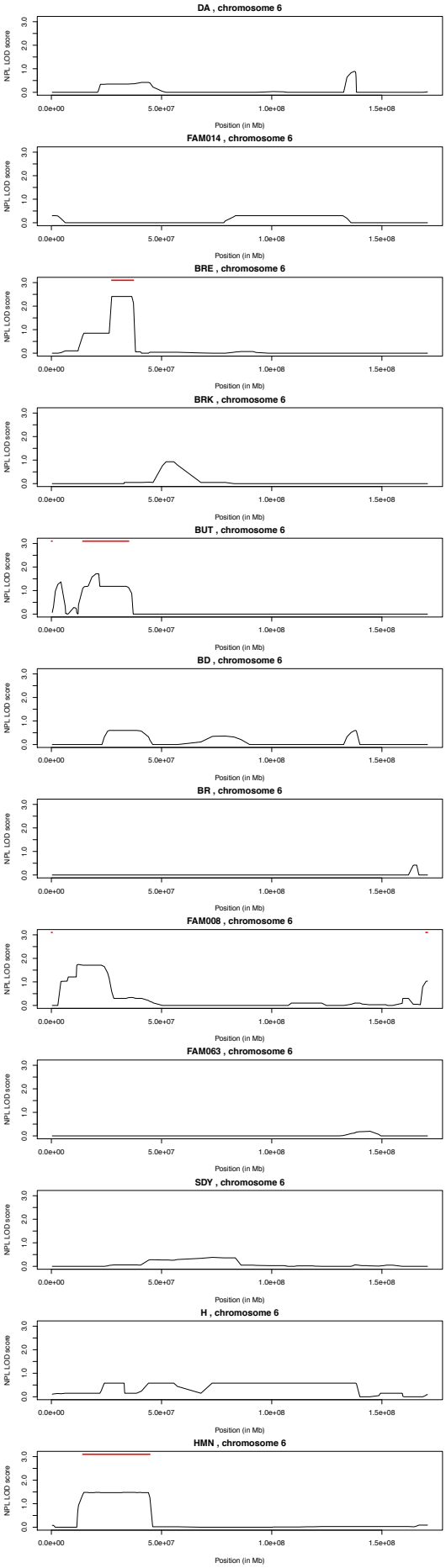

DA , chromosome 1
FAM014 , chromosome 1
BRE , chromosome 1
BRK , chromosome 1
BUT , chromosome 1
BD , chromosome 1
BR , chromosome 1
FAM008 , chromosome 1
FAM063 , chromosome 1
SDY , chromosome 1
H , chromosome 1
HMN , chromosome 1
NPL LOD score
Position (in Mb)
0.0
1.0
2.0
3.0
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
2.5e+08

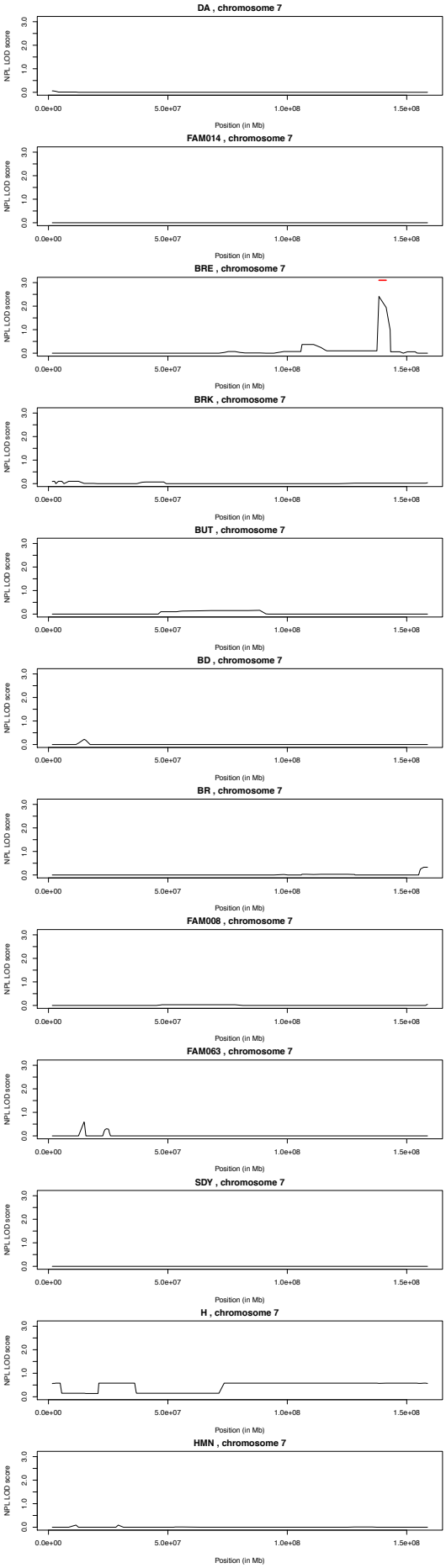

DA , chromosome 2
FAM014 , chromosome 2
BRE , chromosome 2
BRK , chromosome 2
BUT , chromosome 2
BD , chromosome 2
BR , chromosome 2
FAM008 , chromosome 2
FAM063 , chromosome 2
SDY , chromosome 2
H , chromosome 2
HMN , chromosome 2
NPL LOD score
Position (in Mb)
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
2.5e+08
0.0
1.0
2.0
3.0

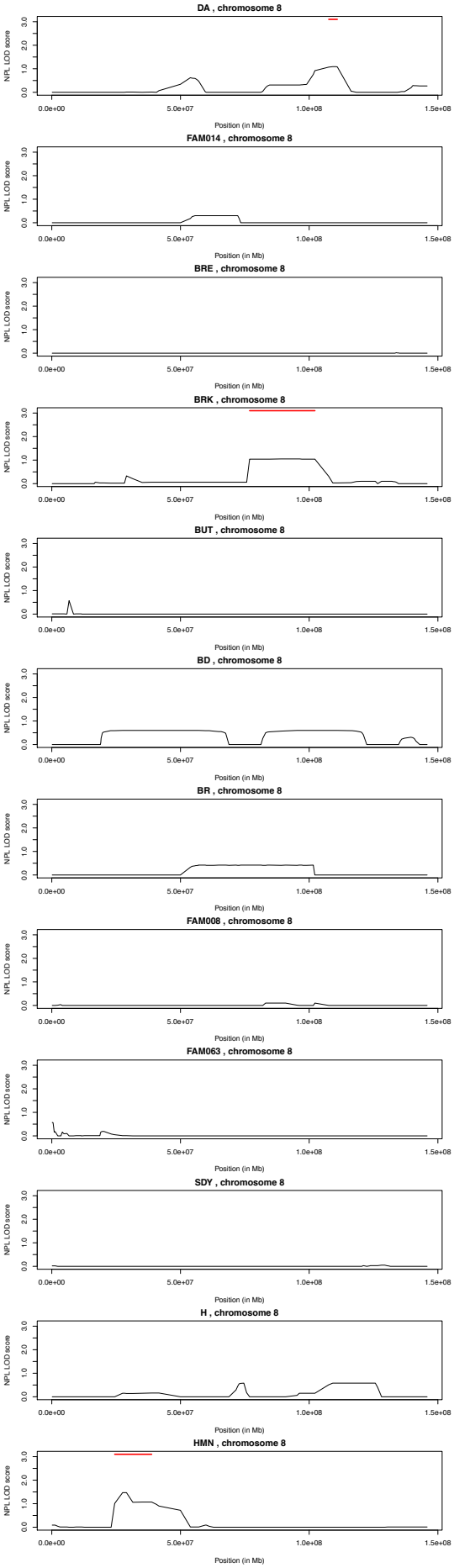

DA , chromosome 3
FAM014 , chromosome 3
BRE , chromosome 3
BRK , chromosome 3
BUT , chromosome 3
BD , chromosome 3
BR , chromosome 3
FAM008 , chromosome 3
FAM063 , chromosome 3
SDY , chromosome 3
H , chromosome 3
HMN , chromosome 3
NPL LOD score
Position (in Mb)
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
0.0
1.0
2.0
3.0

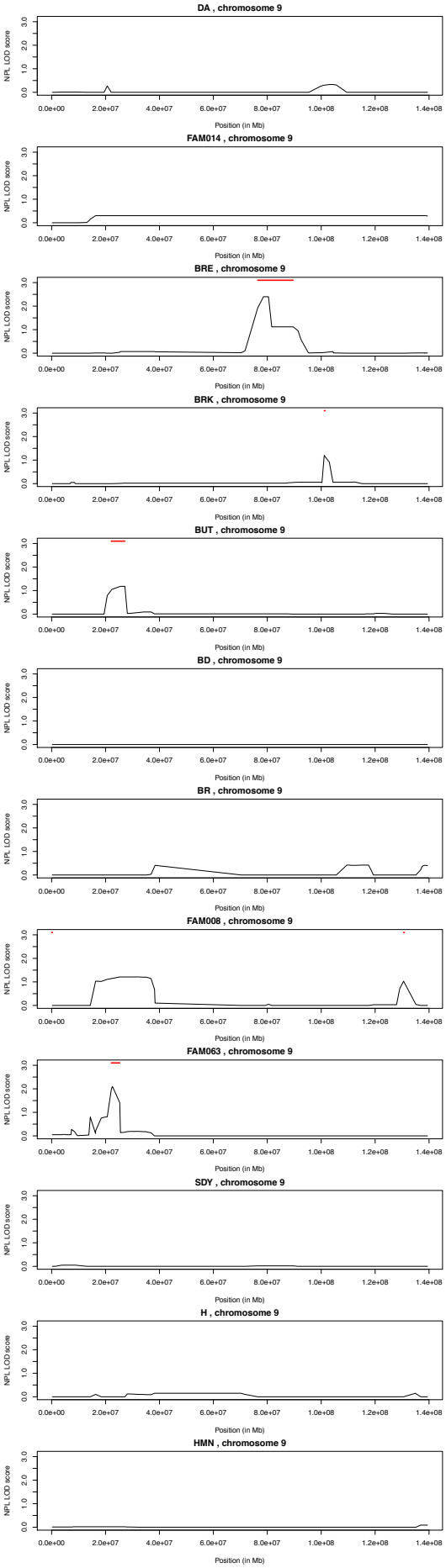

DA , chromosome 4
FAM014 , chromosome 4
BRE , chromosome 4
BRK , chromosome 4
BUT , chromosome 4
BD , chromosome 4
BR , chromosome 4
FAM008 , chromosome 4
FAM063 , chromosome 4
SDY , chromosome 4
H , chromosome 4
HMN , chromosome 4
NPL LOD score
Position (in Mb)
0.0e+00
5.0e+07
1.0e+08
1.5e+08
0.0
1.0
2.0
3.0

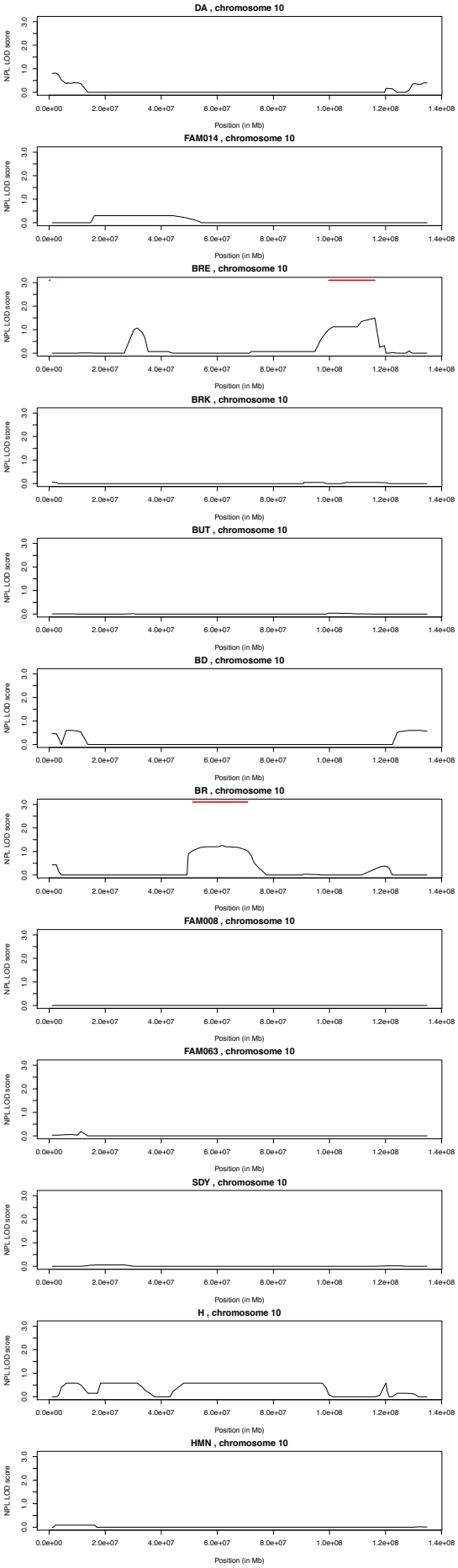

DA , chromosome 5
FAM014 , chromosome 5
BRE , chromosome 5
BRK , chromosome 5
BUT , chromosome 5
BD , chromosome 5
BR , chromosome 5
FAM008 , chromosome 5
FAM063 , chromosome 5
SDY , chromosome 5
H , chromosome 5
HMN , chromosome 5
NPL LOD score
Position (in Mb)
0.0e+00
5.0e+07
1.0e+08
1.5e+08
0.0
1.0
2.0
3.0

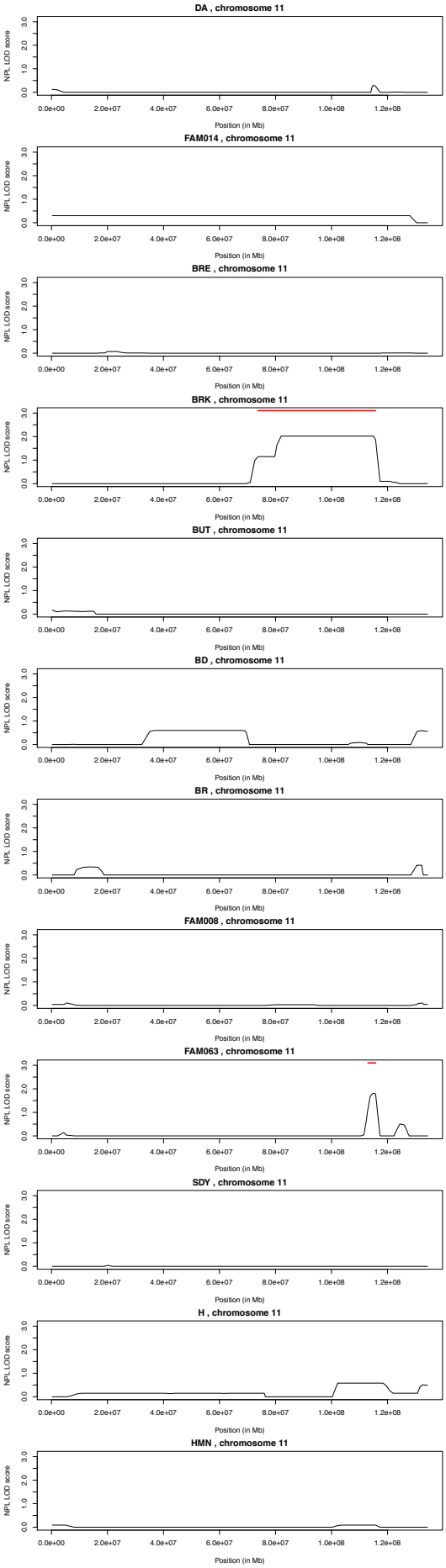

DA , chromosome 6
FAM014 , chromosome 6
BRE , chromosome 6
BRK , chromosome 6
BUT , chromosome 6
BD , chromosome 6
BR , chromosome 6
FAM008 , chromosome 6
FAM063 , chromosome 6
SDY , chromosome 6
H , chromosome 6
HMN , chromosome 6
NPL LOD score
Position (in Mb)
0.0e+00
5.0e+07
1.0e+08
1.5e+08
0.0
1.0
2.0
3.0

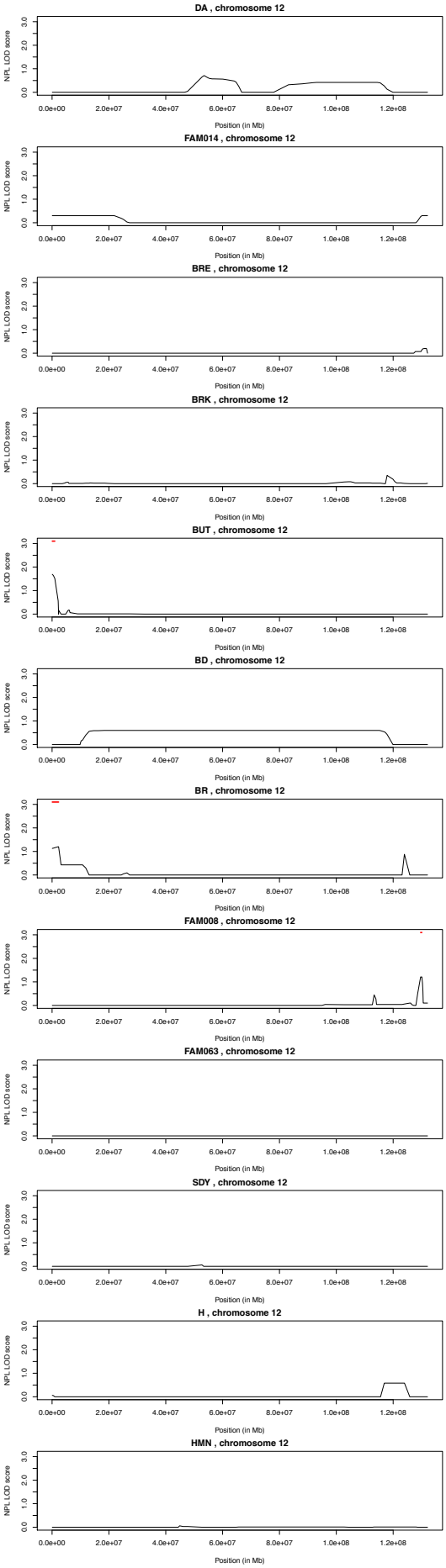

DA , chromosome 7
FAM014 , chromosome 7
BRE , chromosome 7
BRK , chromosome 7
BUT , chromosome 7
BD , chromosome 7
BR , chromosome 7
FAM008 , chromosome 7
FAM063 , chromosome 7
SDY , chromosome 7
H , chromosome 7
HMN , chromosome 7
NPL LOD score
Position (in Mb)
0.0
1.0
2.0
3.0
0.0e+00
5.0e+07
1.0e+08
1.5e+08

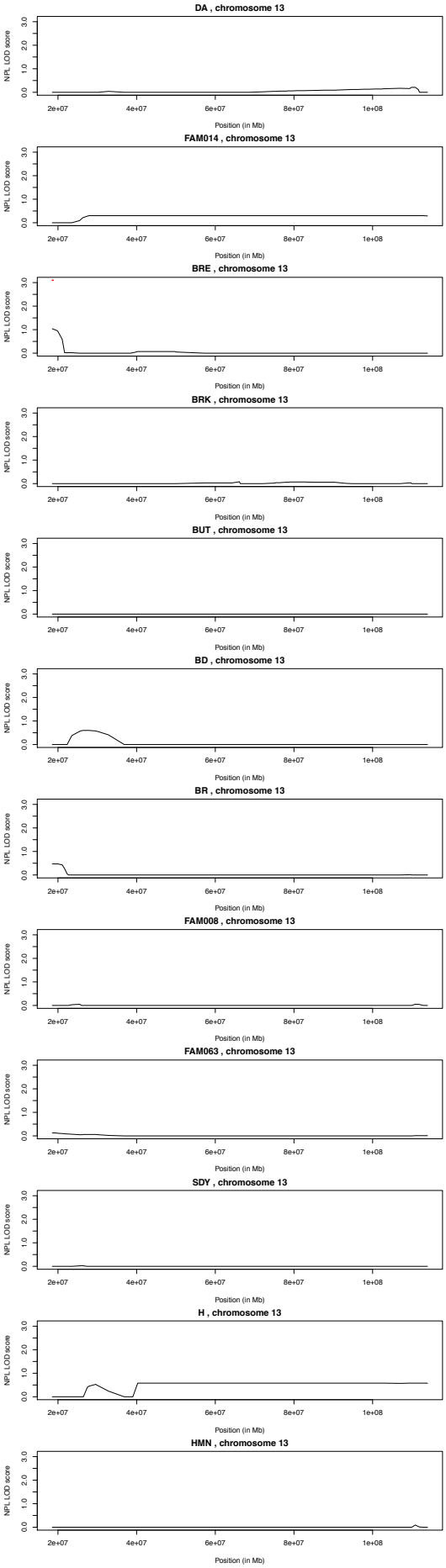

DA , chromosome 8
FAM014 , chromosome 8
BRE , chromosome 8
BRK , chromosome 8
BUT , chromosome 8
BD , chromosome 8
BR , chromosome 8
FAM008 , chromosome 8
FAM063 , chromosome 8
SDY , chromosome 8
H , chromosome 8
HMN , chromosome 8
NPL LOD score
Position (in Mb)
0.0
1.0
2.0
3.0
0.0e+00
5.0e+07
1.0e+08
1.5e+08

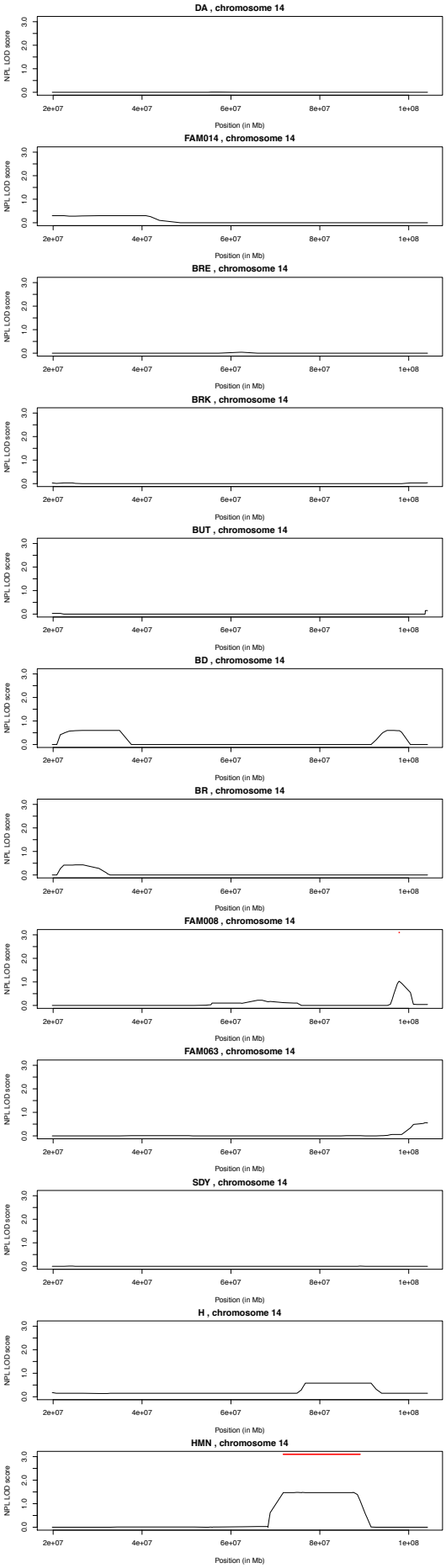

DA , chromosome 9
FAM014 , chromosome 9
BRE , chromosome 9
BRK , chromosome 9
BUT , chromosome 9
BD , chromosome 9
BR , chromosome 9
FAM008 , chromosome 9
FAM063 , chromosome 9
SDY , chromosome 9
H , chromosome 9
HMN , chromosome 9
NPL LOD score
Position (in Mb)
0.0e+00
2.0e+07
4.0e+07
6.0e+07
8.0e+07
1.0e+08
1.2e+08
1.4e+08
0.0
1.0
2.0
3.0

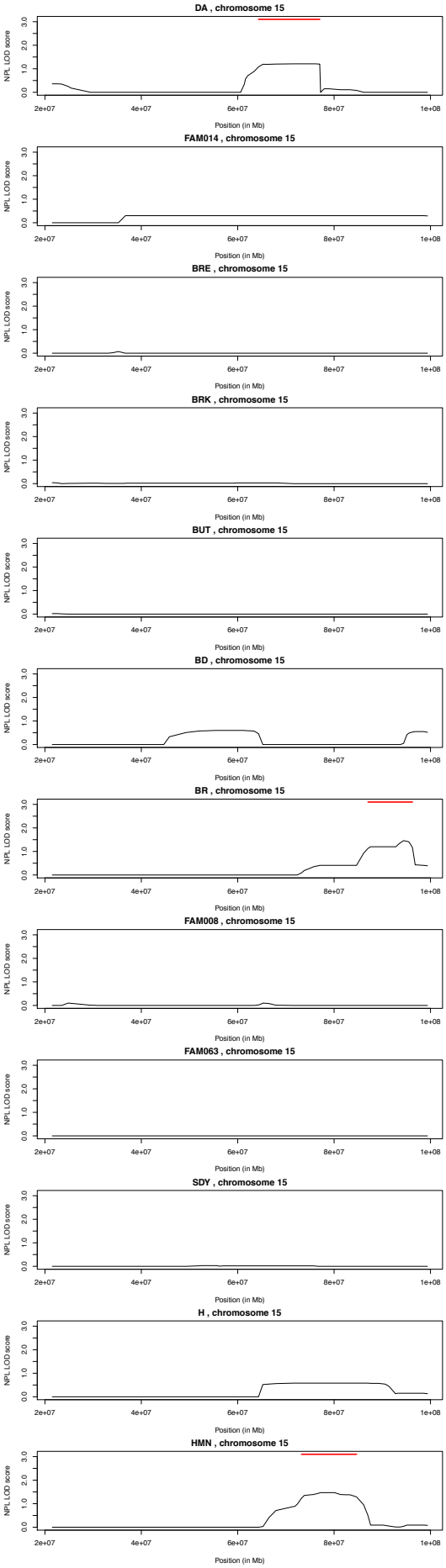

DA , chromosome 10
FAM014 , chromosome 10
BRE , chromosome 10
BRK , chromosome 10
BUT , chromosome 10
BD , chromosome 10
BR , chromosome 10
FAM008 , chromosome 10
FAM063 , chromosome 10
SDY , chromosome 10
H , chromosome 10
HMN , chromosome 10
NPL LOD score
Position (in Mb)
0.0
1.0
2.0
3.0
0.0e+00
2.0e+07
4.0e+07
6.0e+07
8.0e+07
1.0e+08
1.2e+08
1.4e+08

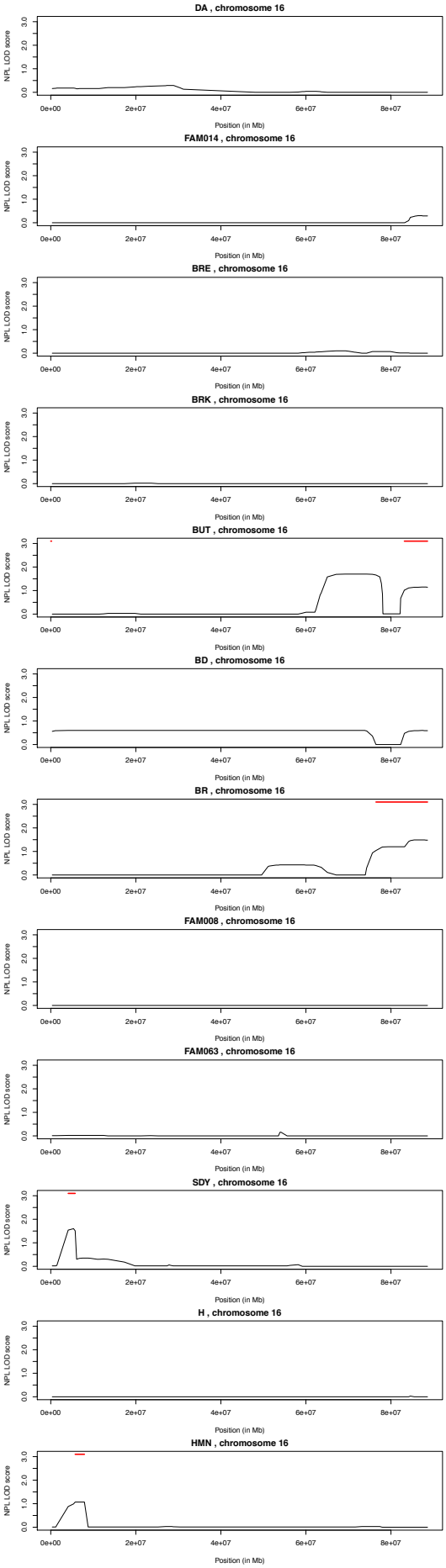

DA , chromosome 11
FAM014 , chromosome 11
BRE , chromosome 11
BRK , chromosome 11
BUT , chromosome 11
BD , chromosome 11
BR , chromosome 11
FAM008 , chromosome 11
FAM063 , chromosome 11
SDY , chromosome 11
H , chromosome 11
HMN , chromosome 11
NPL LOD score
Position (in Mb)
0.0
1.0
2.0
3.0
0.0e+00
2.0e+07
4.0e+07
6.0e+07
8.0e+07
1.0e+08
1.2e+08

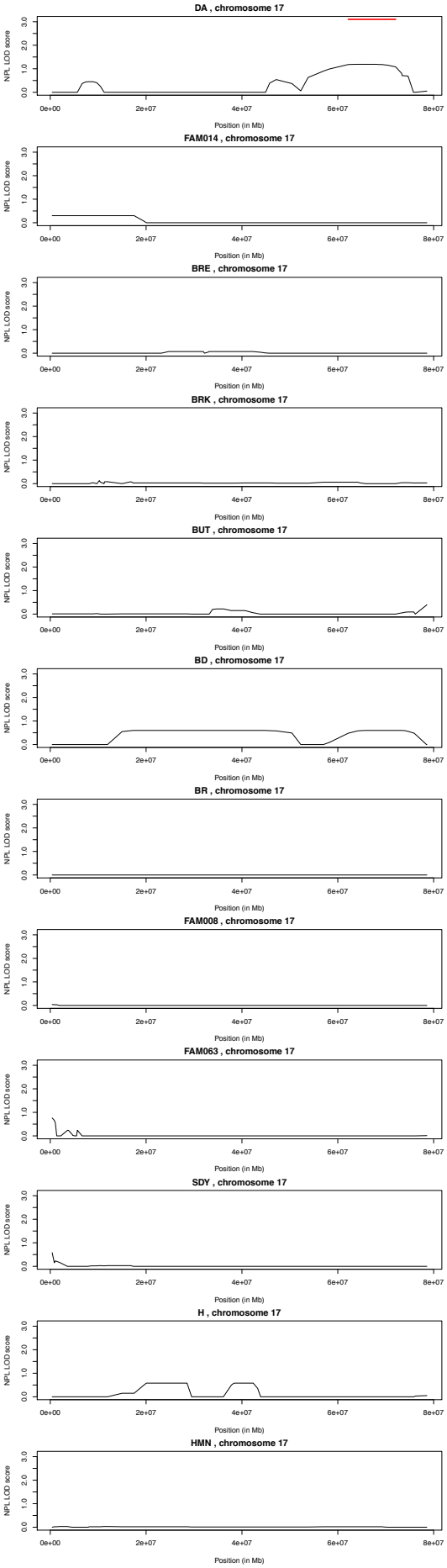

DA , chromosome 12
FAM014 , chromosome 12
BRE , chromosome 12
BRK , chromosome 12
BUT , chromosome 12
BD , chromosome 12
BR , chromosome 12
FAM008 , chromosome 12
FAM063 , chromosome 12
SDY , chromosome 12
H , chromosome 12
HMN , chromosome 12
NPL LOD score
Position (in Mb)
0.0e+00
2.0e+07
4.0e+07
6.0e+07
8.0e+07
1.0e+08
1.2e+08
0.0
1.0
2.0
3.0

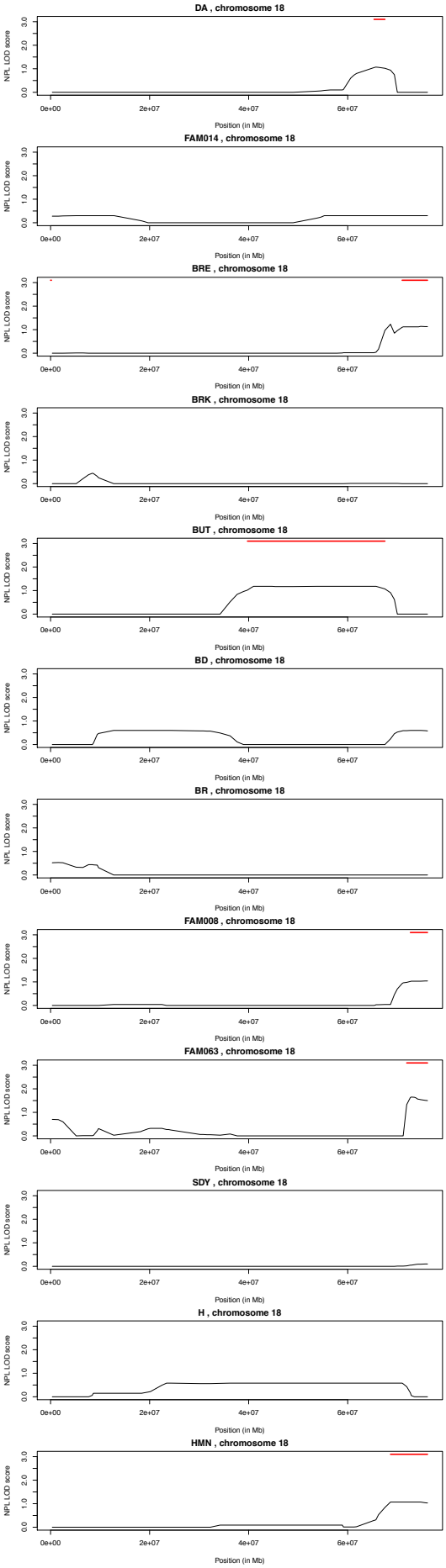

DA , chromosome 13
FAM014 , chromosome 13
BRE , chromosome 13
BRK , chromosome 13
BUT , chromosome 13
BD , chromosome 13
BR , chromosome 13
FAM008 , chromosome 13
FAM063 , chromosome 13
SDY , chromosome 13
H , chromosome 13
HMN , chromosome 13
NPL LOD score
0.0
1.0
2.0
3.0
2e+07
4e+07
6e+07
8e+07
1e+08
Position (in Mb)

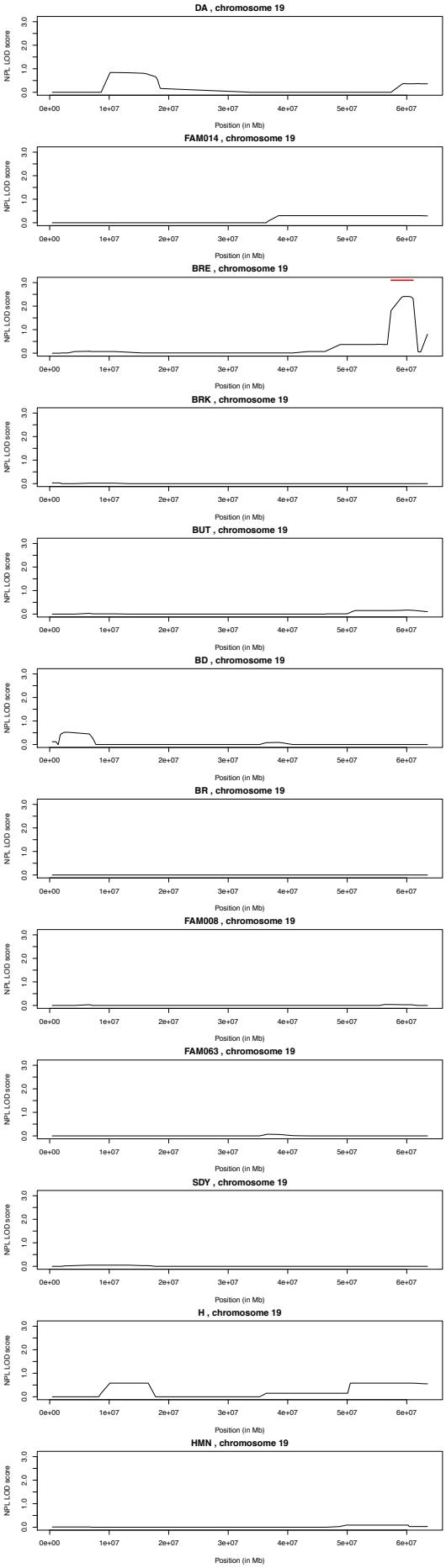

DA , chromosome 14
FAM014 , chromosome 14
BRE , chromosome 14
BRK , chromosome 14
BUT , chromosome 14
BD , chromosome 14
BR , chromosome 14
FAM008 , chromosome 14
FAM063 , chromosome 14
SDY , chromosome 14
H , chromosome 14
HMN , chromosome 14
NPL LOD score
0.0
1.0
2.0
3.0
2e+07
4e+07
6e+07
8e+07
1e+08
Position (in Mb)

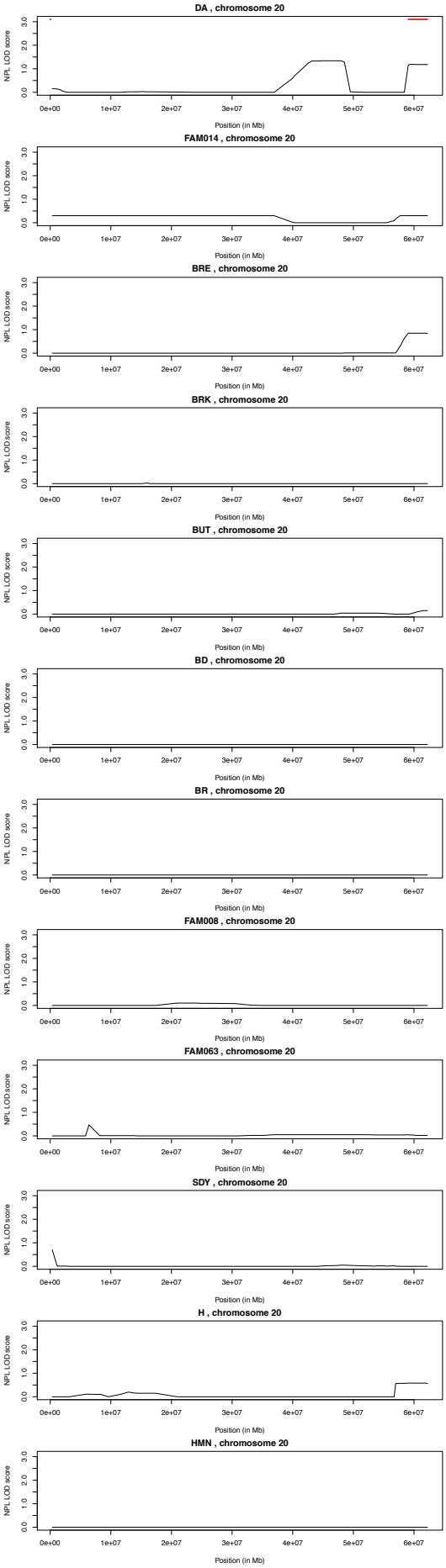

DA , chromosome 15
FAM014 , chromosome 15
BRE , chromosome 15
BRK , chromosome 15
BUT , chromosome 15
BD , chromosome 15
BR , chromosome 15
FAM008 , chromosome 15
FAM063 , chromosome 15
SDY , chromosome 15
H , chromosome 15
HMN , chromosome 15
NPL LOD score
Position (in Mb)
0.0
1.0
2.0
3.0
2e+07
4e+07
6e+07
8e+07
1e+08

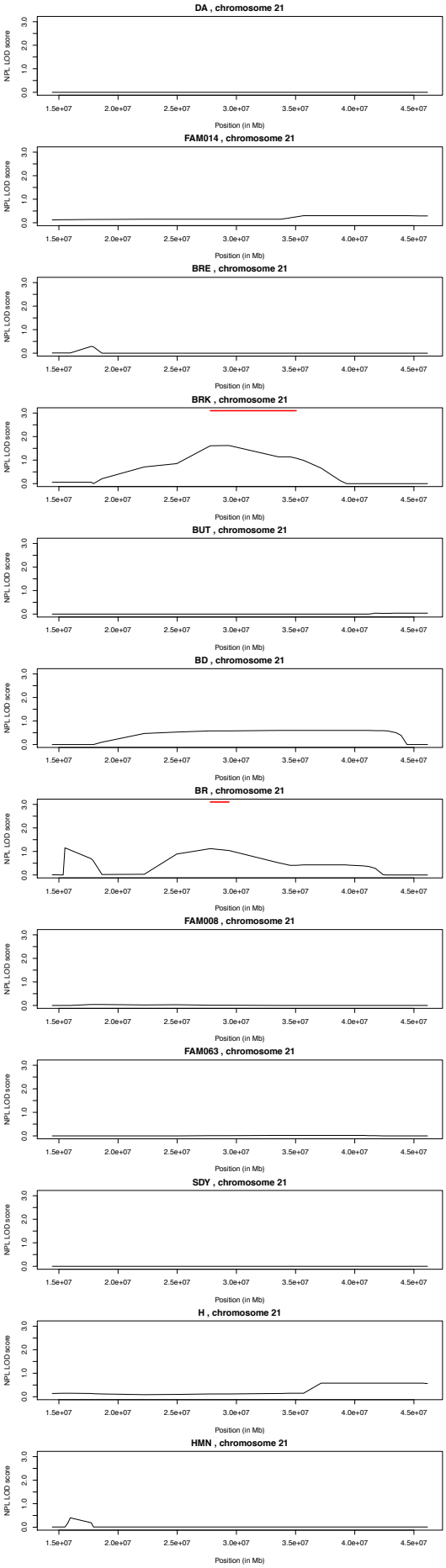

DA , chromosome 16
FAM014 , chromosome 16
BRE , chromosome 16
BRK , chromosome 16
BUT , chromosome 16
BD , chromosome 16
BR , chromosome 16
FAM008 , chromosome 16
FAM063 , chromosome 16
SDY , chromosome 16
H , chromosome 16
HMN , chromosome 16
NPL LOD score
Position (in Mb)
0e+00
2e+07
4e+07
6e+07
8e+07
0.0
1.0
2.0
3.0

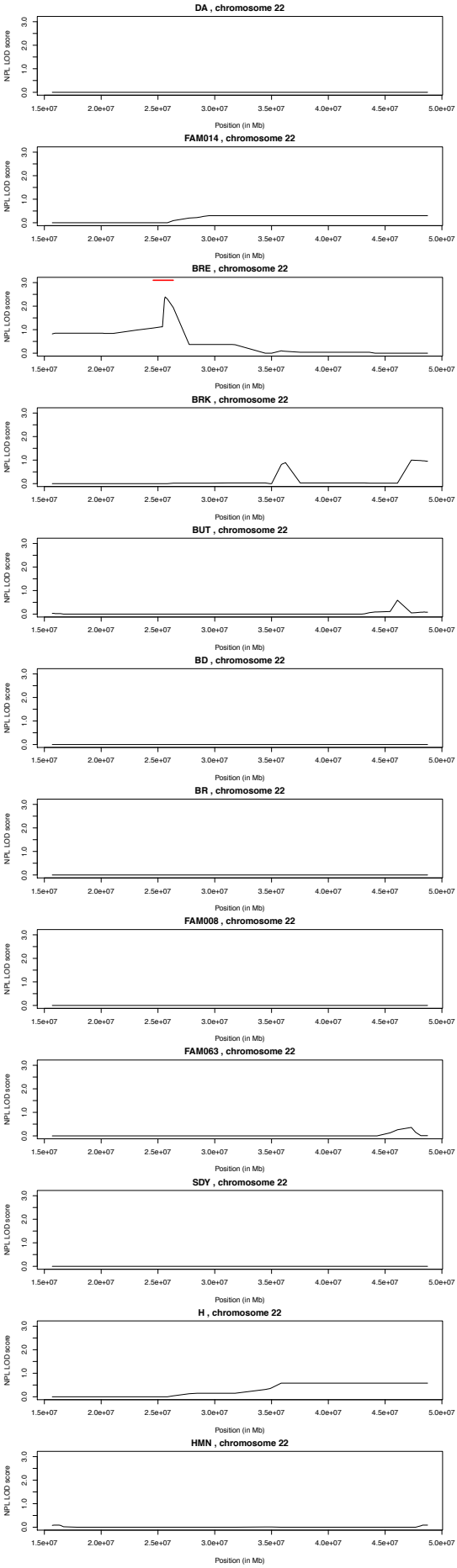

DA , chromosome 17
FAM014 , chromosome 17
BRE , chromosome 17
BRK , chromosome 17
BUT , chromosome 17
BD , chromosome 17
BR , chromosome 17
FAM008 , chromosome 17
FAM063 , chromosome 17
SDY , chromosome 17
H , chromosome 17
HMN , chromosome 17
NPL LOD score
0.0
1.0
2.0
3.0
0e+00
2e+07
4e+07
6e+07
8e+07
Position (in Mb)


DA , chromosome 18
FAM014 , chromosome 18
BRE , chromosome 18
BRK , chromosome 18
BUT , chromosome 18
BD , chromosome 18
BR , chromosome 18
FAM008 , chromosome 18
FAM063 , chromosome 18
SDY , chromosome 18
H , chromosome 18
HMN , chromosome 18
NPL LOD score
Position (in Mb)
0e+00
2e+07
4e+07
6e+07
0.0
1.0
2.0
3.0


DA , chromosome 19
FAM014 , chromosome 19
BRE , chromosome 19
BRK , chromosome 19
BUT , chromosome 19
BD , chromosome 19
BR , chromosome 19
FAM008 , chromosome 19
FAM063 , chromosome 19
SDY , chromosome 19
H , chromosome 19
HMN , chromosome 19
NPL LOD score
0.0
1.0
2.0
3.0
0e+00
1e+07
2e+07
3e+07
4e+07
5e+07
6e+07
Position (in Mb)


DA , chromosome 20
FAM014 , chromosome 20
BRE , chromosome 20
BRK , chromosome 20
BUT , chromosome 20
BD , chromosome 20
BR , chromosome 20
FAM008 , chromosome 20
FAM063 , chromosome 20
SDY , chromosome 20
H , chromosome 20
HMN , chromosome 20
NPL LOD score
Position (in Mb)
0e+00
1e+07
2e+07
3e+07
4e+07
5e+07
6e+07
0.0
1.0
2.0
3.0

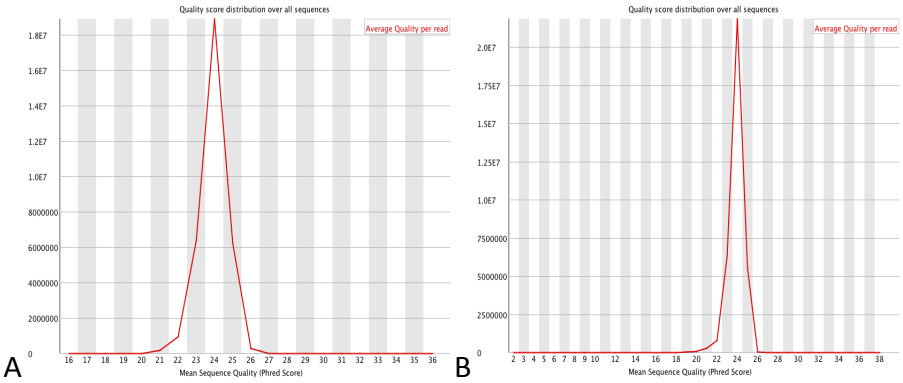

DA , chromosome 21
FAM014 , chromosome 21
BRE , chromosome 21
BRK , chromosome 21
BUT , chromosome 21
BD , chromosome 21
BR , chromosome 21
FAM008 , chromosome 21
FAM063 , chromosome 21
SDY , chromosome 21
H , chromosome 21
HMN , chromosome 21
NPL LOD score
0.0
1.0
2.0
3.0
1.5e+07
2.0e+07
2.5e+07
3.0e+07
3.5e+07
4.0e+07
4.5e+07
Position (in Mb)


DA , chromosome 22
FAM014 , chromosome 22
BRE , chromosome 22
BRK , chromosome 22
BUT , chromosome 22
BD , chromosome 22
BR , chromosome 22
FAM008 , chromosome 22
FAM063 , chromosome 22
SDY , chromosome 22
H , chromosome 22
HMN , chromosome 22
NPL LOD score
Position (in Mb)
0.0
1.0
2.0
3.0
1.5e+07
2.0e+07
2.5e+07
3.0e+07
3.5e+07
4.0e+07
4.5e+07
5.0e+07

Appendix III

Fluidigm pilot study and candidate gene resequencing study results

Figure 1: Quality score distribution for 140bp and 100bp sequencing runs



A: 140bp, 12pM lane; B: 100bp lane. Generated by FastQC software.

Table 1: MiSeq and HiSeq sequencing summaries for 3 x 1536-multiplexed libraries of 506 PCR amplicon sequences; clusters passing filter and density

	Clusters PF %			**Cluster density k/mm2**		
	Lib 01	Lib 02	Lib 03	Lib 01	Lib 02	Lib 03
Concentration	4pM	4pM	4pM	6pM	6pM	5pM
Miseq*	90.4	89.8	88.8	702	685	989
HiSeq 2000**	93.8	93.6	95.4	748	775	640

* 50bp, 10bp index, single end run ** 101bp, 10bp, paired end run

Figure 2: MiSeq index reads for 3 x 1536-multiplexed libraries

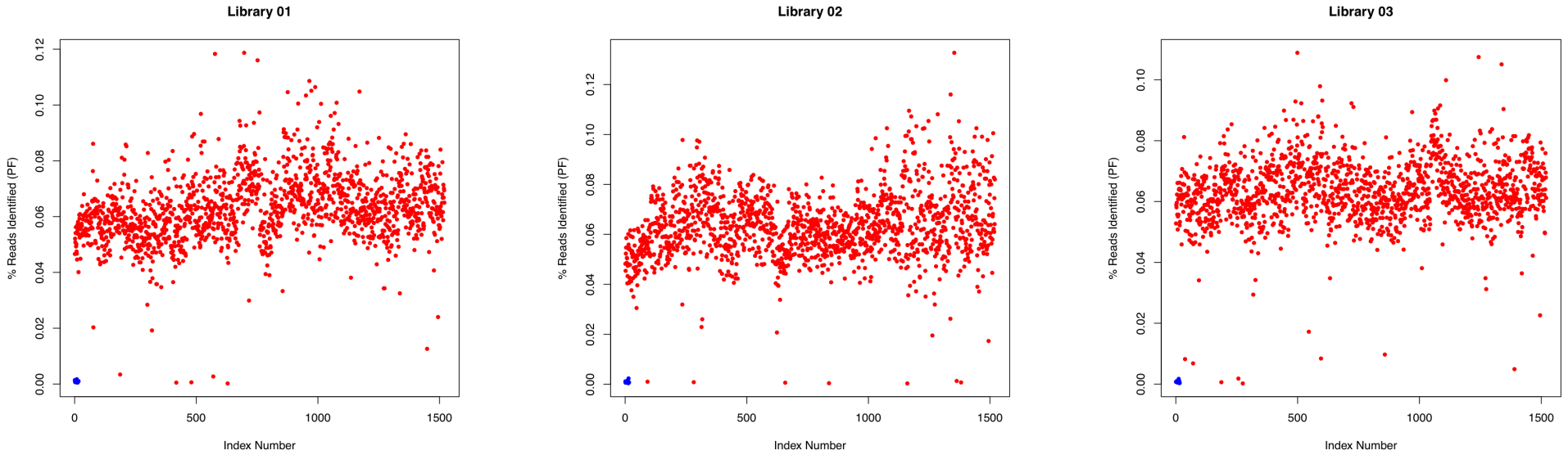
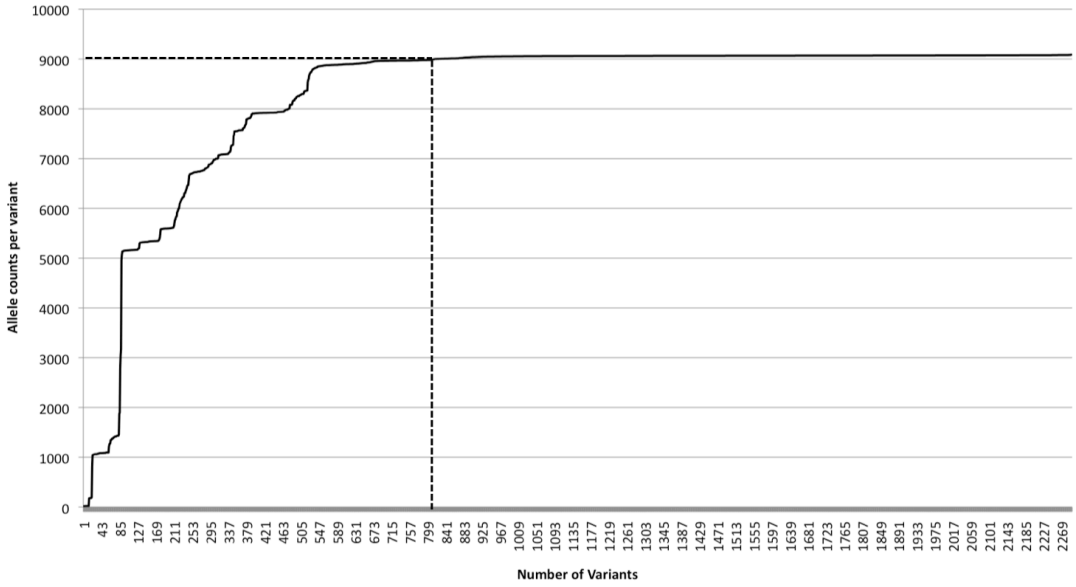
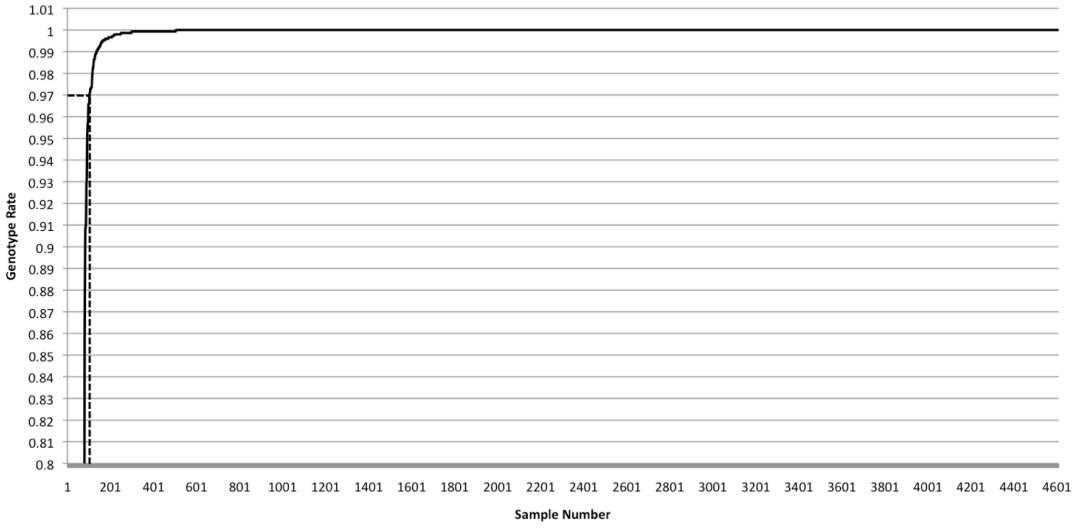


Figure 3: SNP call rate cut off at 2,292 variant sites in 4,608 samples



9000/9216 * 100 = 97.7% call rate applied across all variant sites

Figure 4: Individual genotyping call rate cut off in 4,608 samples



103/4608=2.24% removed based on 97% individual call rate across all variant sites

Figure 5: Alternate allele depth over total allele depth over all heterozygote sites in coding regions of 24 genes

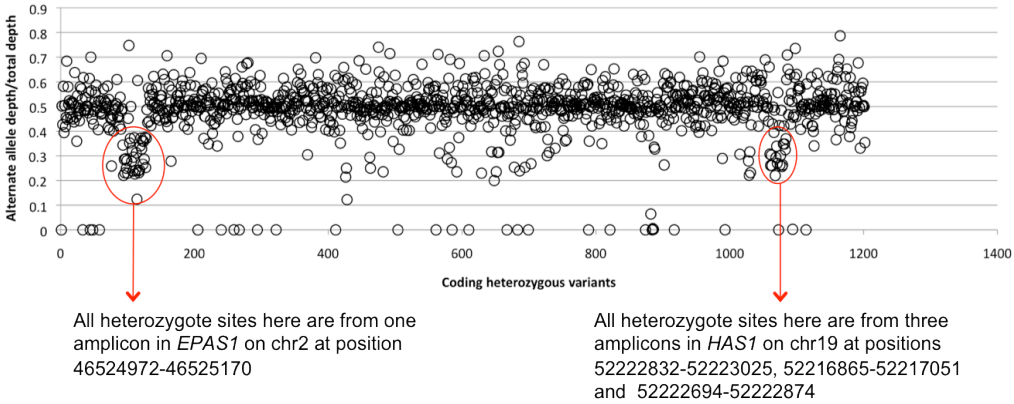
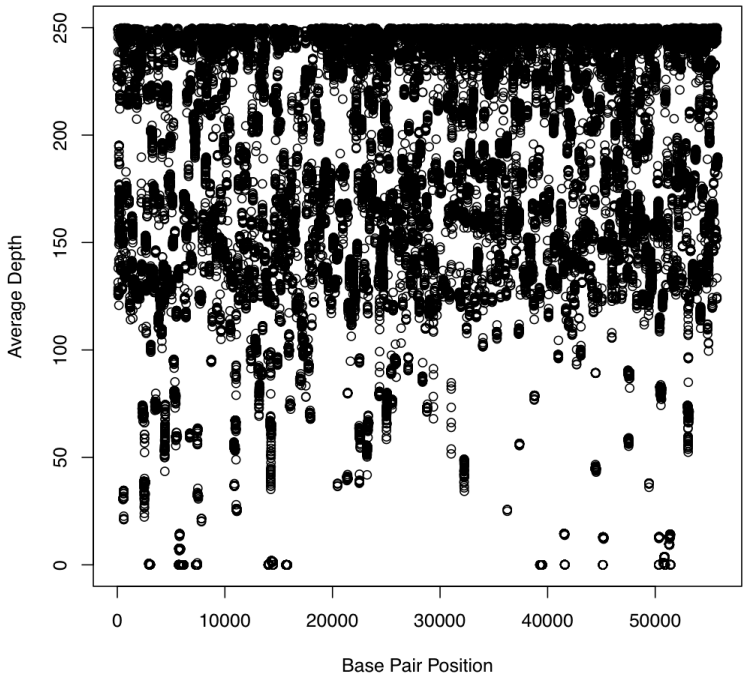


Figure 6: Mean depth per sample



Depth data was produced with a random 100 of 4,478 post quality control samples. GATK settings were minimum base call quality 16, mapping quality >40 and down-sample reads to 250x per sample

Appendix IV

Published papers

Hunt, K. A., **V. Mistry**, et al. (2013). "Negligible impact of rare autoimmune locus coding-region variants on missing heritability." Nature.

Hunt, K. A., D. J. Smyth, T. Balschun, M. Ban, **V. Mistry** et al. (2012). "Rare and functional SIAE variants are not associated with autoimmune disease risk in up to 66,924 individuals of European ancestry." Nature Genetics 44(1): 3-5.

Mistry, V. and D. van Heel (2011). "Molecular Genetics of Coeliac Disease." eLS.

Trynka, G., K. A. Hunt, N.A. Bockett, J. Romanos, **V. Mistry**. (2011). "Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease." Nature Genetics 43(12): 1193-1201.

Dubois, P.C., G. Trynka, L. Franke, K.A. Hunt, J. Romanos, A. Curtotti, A Zhernakova, G.A. Heap, R. Adany, A. Aromaa, M.T. Bardella, L.H. van den Berg, N.A. Bockett, E.G. de la Concha, B. Dema, R.S. Fehrmann, M. Fernandez-Arquero, S. Fiatal, E. Grandone, P.M. Green, H.J. Groen, R.G. William, R.H. Houwen, S.E. Hunt, K. Kaukinen, D. Kelleher, I. Korponay-Szabo, K. Kurppa, P. MacMathuna, M. Maki, M.C. Mazzilli, O.T. McCann, M.L. Mearin, C.A. Mein, **V. Mistry**, et al. (2010). “Multiple common variants for celiac disease influencing immune gene expression.” Nature Genetics 42(4): 295-302.

All publications pertaining to the work in this thesis are below.

LETTER

doi:10.1038/nature12170

Negligible impact of rare autoimmune-locus coding-region variants on missing heritability

Karen A. Hunt[1], Vanisha Mistry[1], Nicholas A. Bockett[1], Tariq Ahmad[2], Maria Ban[3], Jonathan N. Barker[4], Jeffrey C. Barrett[5], Hannah Blackburn[5], Oliver Brand[6], Oliver Burren[7], Francesca Capon[4], Alastair Compston[3], Stephen C. L. Gough[6], Luke Jostins[8], Yong Kong[9], James C. Lee[10], Monkol Lek[11], Daniel G. MacArthur[11], John C. Mansfield[12], Christopher G. Mathew[4], Charles A. Mein[13], Muddassar Mirza[4], Sarah Nutland[7], Suna Onengut-Gumuscu[14], Efterpi Papouli[4], Miles Parkes[10], Stephen S. Rich[14], Steven Sawcer[3], Jack Satsangi[15], Matthew J. Simmonds[6], Richard C. Trembath[16], Neil M. Walker[7], Eva Wozniak[13], John A. Todd[7], Michael A. Simpson[4], Vincent Plagnol[17] & David A. van Heel[1]

Genome-wide association studies (GWAS) have identified common variants of modest-effect size at hundreds of loci for common autoimmune diseases; however, a substantial fraction of heritability remains unexplained, to which rare variants may contribute[1,2]. To discover rare variants and test them for association with a phenotype, most studies re-sequence a small initial sample size and then genotype the discovered variants in a larger sample set[3–5]. This approach fails to analyse a large fraction of the rare variants present in the entire sample set. Here we perform simultaneous amplicon-sequencing-based variant discovery and genotyping for coding exons of 25 GWAS risk genes in 41,911 UK residents of white European origin, comprising 24,892 subjects with six autoimmune disease phenotypes and 17,019 controls, and show that rare coding-region variants at known loci have a negligible role in common autoimmune disease susceptibility. These results do not support the rare-variant synthetic genome-wide-association hypothesis[6] (in which unobserved rare causal variants lead to association detected at common tag variants). Many known autoimmune disease risk loci contain multiple, independently associated, common and low-frequency variants, and so genes at these loci are a priori stronger candidates for harbouring rare coding-region variants than other genes. Our data indicate that the missing heritability for common autoimmune diseases may not be attributable to the rare coding-region variant portion of the allelic spectrum, but perhaps, as others have proposed, may be a result of many common-variant loci of weak effect[7–10].

Recent large-scale human sequencing studies have revealed an abundance of rare variants (which we define as minor allele frequency (MAF) $< 0.5\%$) and shown that these are geographically localized and are more likely to have deleterious functional consequences[11,12]. In the largest sample size studied to date[12], 202 genes in 14,002 people were re-sequenced, and ~95% of exonic variants identified were found to be rare, with 74% observed in only one or two subjects. More broadly, across ~15,000 genes, similar findings were observed in recent exome-sequencing studies of 2,440 and 6,515 subjects[13,14]. Importantly, these studies demonstrate that even if we had reference variation databases from a million subjects, most of the rare-variant allelic spectrum of any given sample set (for example, a case–control cohort) will be unique and only identifiable by direct re-sequencing of the entire sample set.

There are only a handful of published examples of rare coding-region variants associated with common autoimmune diseases (although many examples in familial/Mendelian immune-mediated diseases). Coding-region variants in *IFIH1* associated with type 1 diabetes (MAF in controls = 0.67–2.2%)[3], *TYK2* with multiple autoimmune diseases[15] and *IL23R* with inflammatory bowel disease[5], for example, are low frequency (which we define as MAF = 0.5–5%) rather than particularly rare. In other examples, the existing evidence for association, and/or the effect sizes, are relatively weak (for example, *CARD14* and psoriasis[16], *IL2RA* and *IL2RB* and rheumatoid arthritis[17]). The association of rare coding-region variants of *NOD2* (also known as *CARD15*) in Crohn's disease probably provides the best example, albeit three low-frequency variants comprise over 80% of all the disease-causing mutations[18]. Most of the studies also lose power (especially for tests in which multiple rare variants are pooled into a single analysis, for example by gene) by initially sequencing only a small sample subset rather than testing the entire rare-variant content of a large case–control sample set. We sought to improve on these methods by performing highly multiplexed sequencing of sufficiently high quality to enable direct genotyping in the entirety of a large autoimmune disease case–control collection.

We selected subjects from a single population—individuals of white Northern-European ethnicity living in the UK (Methods)—to minimize any effects of population stratification. We selected to re-sequence all RefSeq exons for 25 genes from 20 GWAS-identified risk loci showing overlap between six common autoimmune disease phenotypes (autoimmune thyroid disease, coeliac disease, Crohn's disease, psoriasis, multiple sclerosis and type 1 diabetes). All genes studied were from risk loci for at least two phenotypes, all genes had known immune system function, 18 out of 20 loci had either a single candidate immune gene or all immune genes at a locus were selected (the remaining two loci had partial transcripts of another immune gene within the 0.1 centimorgan (cM) linkage disequilibrium block), and all genes and loci were densely genotyped on the Illumina ImmunoChip (Supplementary Table 1)[19]. We attempted high-throughput sequencing of 52,224 samples (including positive and negative controls, and repeats). We performed extensive quality control on both samples and variant calls (Methods). The final data set comprised 41,911 phenotyped individuals (autoimmune disease cases and controls), with ImmunoChip

[1]Blizard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London E1 2AT, UK. [2]Peninsula College of Medicine and Dentistry, Barrack Road, Exeter EX2 5DW, UK. [3]University of Cambridge, Department of Clinical Neurosciences, Addenbrooke's Hospital, Cambridge CB2 0QQ, UK. [4]Division of Genetics and Molecular Medicine, King's College London School of Medicine, 8th Floor Tower Wing, Guy's Hospital, London SE1 9RT, UK. [5]Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK. [6]Oxford Centre for Diabetes Endocrinology and Metabolism, University of Oxford, Oxford OX3 7LJ, UK. [7]Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory, Department of Medical Genetics, Cambridge Institute for Medical Research, University of Cambridge, Cambridge CB2 0XY, UK. [8]Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK. [9]Department of Molecular Biophysics and Biochemistry, W.M. Keck Foundation Biotechnology Resource Laboratory, Yale University, New Haven, Connecticut 06510, USA. [10]Department of Medicine, University of Cambridge School of Clinical Medicine, Addenbrooke's Hospital, Cambridge CB2 0QQ, UK. [11]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [12]Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne NE1 3BZ, UK. [13]Genome Centre, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, John Vane Science Centre, Charterhouse Square, London EC1M 6BQ, UK. [14]Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia 22908-0717, USA. [15]Gastrointestinal Unit, Molecular Medicine Centre, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, UK. [16]Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London E1 2AT, UK. [17]University College London Genetics Institute, Gower Street, London WC1E 6BT, UK.

Table 1 | Variant types in protein-coding regions of 25 genes in 41,911 phenotyped individuals

Variant type	All variants	Rare (MAF < 0.5%)*	Novel†
Nonsynonymous SNV	1,792	1,758	1,379
Splicing SNV	86	85	65
Stopgain SNV	47	47	42
Synonymous SNV	1,024	972	674
Frameshift indels	31	31	31
Nonframeshift indels	10	10	10
Total variants	2,990	2,903	2,201
Singleton	1,602	1,598	1,411
Doubleton	470	468	378

Numbers shown are after quality-control steps. Annotation performed with GENCODE V14 gene definitions. Triallelic ($n = 124$) and quadrallelic ($n = 3$) sites (combined SNVs and indels) are shown as multiple separate variants with the appropriate annotation for each non-reference allele.
* MAF in 17,019 sequenced controls.
† Not seen in dbSNP137, or 1000 Genomes Project (April 2012 release), or NHLBI (data release ESP6500SI, with 6,503 individuals).

array genotypes available for 32,806 of these individuals (Supplementary Table 2). We discovered 4,377 variant sites across all amplicons, and the genotype call rate was 99.9989% (reference homozygote as well as non-reference genotypes) across 41,911 individuals. Of these, 2,990 variants were in protein-coding regions (including exon splice sites) of the 25 genes (Table 1 and Supplementary Table 3); 97.1% of which are rare (MAF in 17,019 controls, <0.5%); 73.6% are novel when compared with current published data sets (dbSNP137, 1000 Genomes Project, National Heart, Lung, and Blood Institute (NHLBI)) containing >6,000 individuals and 67.3% are novel compared to an unpublished data set of 25,994 exome-sequenced individuals (D. G. MacArthur, personal communication); and 68.9% were only seen in one (singleton) or two (doubleton) individuals. These proportions of novel, and rare, variants are similar to recent data from other large re-sequencing studies[12].

Our very high coverage data (99.8% of 183.4 million (site X sample) genotype calls had a read depth of ≥40 and 96.6% had a read depth of >100; Supplementary Fig. 1) enabled stringent data filtering on call rate per sample, per variant site, and other criteria (Methods). To confirm data quality, we performed further experiments and analyses as follows: (1) we genotyped one control sample 296 times (on different 48-sample microfluidic chips), and the genotype call error rate was two non-consensus genotype calls of 1,295,581 called genotypes (0.00015%); (2) 32,806 out of 41,911 subjects also had dense ImmunoChip genotyping data at the 25 genes, and genotype concordance at 91 variant sites genotyped on both platforms was 99.994%; (3) transition/transversion (Ti/Tv) rates, a quality-control measure based on expected human mutation types, were 2.434 at coding-region variants (2.427 at singletons), 2.44 at rare (MAF < 0.5%) variants (2.437 at singletons) and 2.275 at novel variants (2.273 at singletons) (definitions in Table 1); (4) we selected all (35) nonsense single nucleotide variants (SNVs) and all (39) frameshift insertions/deletions (indels) in the ImmunoChip-genotyped samples for Sanger sequencing: two variants failed assay/PCR (polymerase chain reaction) design and there was one false-positive SNV and one false-positive indel (overall false-positive rate = 2.8%). All 70 validated SNVs and indels had the same alleles in high-throughput and Sanger-sequencing

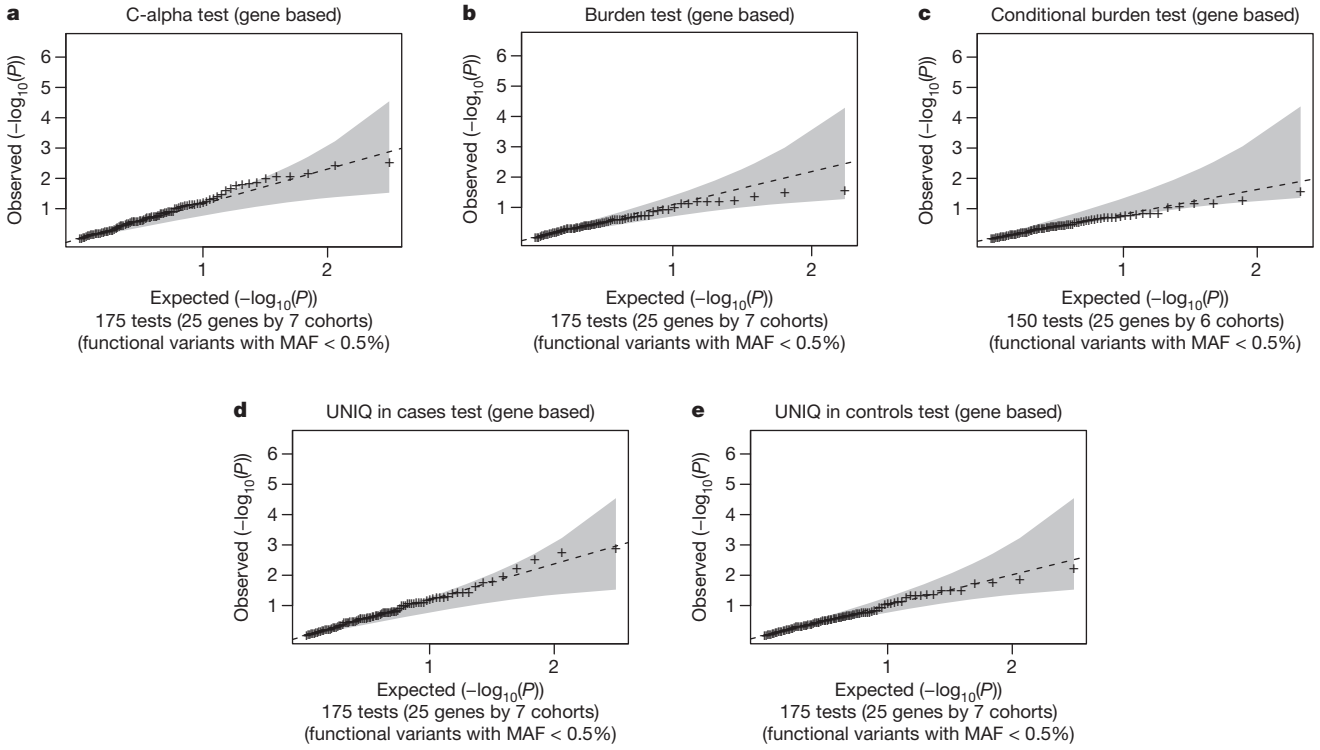


Figure 1 | Association analyses of discovered rare functional variants in autoimmune diseases. We define rare functional variants as MAF < 0.5% in 17,019 controls and predicted nonsynonymous, premature-stop or splice-site annotation. Quantile–quantile plots compare observed versus expected test-statistic distributions, with shading indicating 99% confidence intervals. Full results are available in Supplementary Data. Each of six individual diseases, and all autoimmune diseases combined, were tested as phenotypes. **a**, Gene-based C-alpha test (25 genes by 7 phenotypes, $n = 41,911$ subjects) allowing for both risk and protective effects for rare functional variants. Singleton variants pooled into a single binomial count per phenotype. **b**, Gene-based burden tests (25 genes by 7 phenotypes, $n = 41,911$ subjects) comparing summed allele counts for rare functional variants in cases versus controls with Fisher's exact test. **c**, Conditional gene-based burden test (25 genes by 6 phenotypes, $n = 32,806$ subjects): rare functional-variant allele counts are summed for each individual per gene and introduced in a logistic regression, including ImmunoChip covariates for multiple independent top (common) variant signals selected on the basis of a stepwise regression (down to $P > 10^{-4}$). The psoriasis phenotype was not tested as most samples do not have ImmunoChip data. **d**, Count of case-unique rare alleles (UNIQ) tests (25 genes by 7 phenotypes, $n = 41,911$ subjects): compares the number of rare functional variants only observed in cases with the distribution of this value upon random permutation (10,000 times) of the phenotypes. **e**, Count of control-unique rare alleles (UNIQ) tests: same as **d** but for rare functional variants uniquely observed in controls.

assays; (5) proportions of rare, and of known, variants were similar to those found by other large sequencing studies, and we identified no common or low-frequency novel variant sites.

We first attempted to identify any low-frequency or rare variants of larger effect. We performed for each coding-region variant and each of seven phenotypes (including all autoimmune disease cases combined) a single-variant association analysis. Only previously reported loci were observed with common variants (MAF > 5%), as expected. We identified three low-frequency (MAF = 0.5–5%) and rare (MAF in 17,019 controls = <0.5%) exonic variants with single SNP association $P < 10^{-4}$ (chosen as a partial Bonferroni multiple testing correction for 25 genes and 7 phenotypes, but not correcting for all variants per gene) (Supplementary Table 4 and Supplementary Data). We next analysed low-frequency and rare exonic variants, conditioning on common-variant non-coding signals at each locus, and observed no additional association signals (Supplementary Data). An association between type 1 diabetes and the low-frequency *UBASH3A* SNP rs17114930 was observed, but conditional regression analysis showed this signal to be secondary to a stronger common-frequency variant/haplotype previously identified by GWAS[20]. We identified novel low-frequency (nearly 'common' as MAF in 17,019 controls = 4.97%) *NCF2* coding-region variant associations with coeliac disease at two SNPs (rs17849502, nonsynonymous; rs17849501, synonymous; in almost complete linkage disequilibrium $r^2 = 0.992$). Both variants were present on the Illumina ImmunoChip, but just failed quality-control criteria in our previous coeliac disease study owing to missing data[19]. We replicated the UK findings in 4,313 coeliac cases and 3,954 controls (European samples, Methods; rs17849502 $P = 4.46 \times 10^{-5}$ (Cochran–Mantel–Haenszel test), odds ratio 1.35 (95% CI = 1.17–1.55)). Logistic regression analysis conditioning on rs17849502 in the UK re-sequencing data set revealed no further single-variant coeliac disease association signals below $P < 10^{-4}$. *NCF2* is a component of the neutrophil NADPH oxidase respiratory burst complex. Different disease-causing mutations cause the recessive Mendelian phenotype chronic granulomatous disease. The rs17849502/H389Q variant is also associated with the autoimmune disease systemic lupus erythematosus[21]. Functional studies have shown that the minor allele of rs17849502/H389Q reduces the binding efficiency of NCF2 to the guanine nucleotide-exchange factor VAV1 (ref. 21). These data now implicate a disease mechanism of impaired neutrophil function in coeliac disease, a condition previously thought to be of predominantly B- and T-cell-mediated immunopathogenesis, and where neutrophils may have a role in regulating adaptive immunity[22].

We noted that even with ~7,000 cases and ~17,000 controls the power to detect association signals using single-variant tests for variants (MAF < 0.5%) of modest effect (for example, odds ratio < 3) is limited (Supplementary Fig. 2) and therefore we performed gene-based pooled-variant association tests to better detect the combined effect of multiple variants. We defined coding-region variants as functional candidates if the variants were rare (MAF in 17,019 controls = <0.5%) and predicted to be of potential functional impact (nonsynonymous, premature stop, splice-site altering; see Methods). We pooled variants (by gene) in analyses to detect different scenarios (Fig. 1 and Supplementary Data), including the C-alpha test, which can detect a combination of risk and protective variants; burden tests to detect either an excess of risk variants in cases or protective variants in controls; a modified version of the burden test using conditional regression and common-variant non-coding signals at a locus as covariates; a test to detect an excess of rare variants seen uniquely in cases (the case or control unique tests being particularly suitable for the study of the large numbers of singleton and doubleton variants we observe); and a test to detect an excess of rare variants seen uniquely in controls. The distribution of association statistics for all five pooled gene tests across each of the six or seven phenotypes tested was consistent with the global null of no association.

On the basis of these results, in the largest (to the best of our knowledge) human disease sample sequencing study to date, we find little support for a significant impact of rare coding-region variants in known risk genes for the autoimmune disease phenotypes tested. Our data provide little stimulus in support of large-scale whole-exome sequencing projects in common autoimmune diseases. Using average genetic-effect estimates from our data (Methods), over all loci and phenotypes we have tested, we estimate that rare variants contribute to less than 3% of the heritability explained by common variants at these known risk loci[23].

METHODS SUMMARY

Sequencing. DNA (corresponding to exonic sequence of 25 autoimmune disease risk genes) was PCR-amplified in a multiplexed microfluidics assay (Fluidigm Access Array). PCR amplicons from a sample were pooled, and barcoded with one of 1,536 unique ten-base-pair sequences. Libraries of 1,536 samples were sequenced on Illumina HiSeq instruments. Reads were aligned to the GRCh37 human reference and SNVs and small indels called. Samples and called variants were extensively filtered on the basis of call rate and other criteria. Selected variants were validated by Sanger dideoxy sequencing. Genotype data from Illumina ImmunoChip array-based genotyping was merged with Fluidigm sequencing-based genotypes.

Statistical analysis. Statistical analysis was performed in R, and using PLINK/SEQ software.

Full Methods and any associated references are available in the online version of the paper.

Received 27 February; accepted 8 April 2013.
Published online 22 May 2013.

1. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).
2. Gibson, G. Rare and common variants: twenty arguments. *Nature Rev. Genet.* **13,** 135–145 (2012).
3. Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J. A. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* **324,** 387–389 (2009).
4. Rivas, M. A. *et al.* Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nature Genet.* **43,** 1066–1073 (2011).
5. Momozawa, Y. *et al.* Resequencing of positional candidates identifies low frequency *IL23R* coding variants protecting against inflammatory bowel disease. *Nature Genet.* **43,** 43–47 (2011).
6. Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D. B. Rare variants create synthetic genome-wide associations. *PLoS Biol.* **8,** e1000294 (2010).
7. Bloom, J. S., Ehrenreich, I. M., Loo, W. T., Lite, T. L. & Kruglyak, L. Finding the sources of missing heritability in a yeast cross. *Nature* **494,** 234–237 (2013).
8. Stahl, E. A. *et al.* Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nature Genet.* **44,** 483–489 (2012).
9. Park, J. H. *et al.* Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature Genet.* **42,** 570–575 (2010).
10. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genet.* **42,** 565–569 (2010).
11. Coventry, A. *et al.* Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature Commun.* **1,** 131 (2010).
12. Nelson, M. R. *et al.* An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337,** 100–104 (2012).
13. Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337,** 64–69 (2012).
14. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493,** 216–220 (2013).
15. Strange, A. *et al.* A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between *HLA-C* and *ERAP1*. *Nature Genet.* **42,** 985–990 (2010).
16. Jordan, C. T. *et al.* Rare and common variants in *CARD14*, encoding an epidermal regulator of NF-kappaB, in psoriasis. *Am. J. Hum. Genet.* **90,** 796–808 (2012).
17. Diogo, D. *et al.* Rare, low-frequency, and common variants in the protein-coding sequence of biological candidate genes from GWASs contribute to risk of rheumatoid arthritis. *Am. J. Hum. Genet.* **92,** 15–27 (2013).
18. Lesage, S. *et al. CARD15/NOD2* mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. *Am. J. Hum. Genet.* **70,** 845–857 (2002).
19. Trynka, G. *et al.* Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nature Genet.* **43,** 1193–1201 (2011).
20. Barrett, J. C. *et al.* Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature Genet.* **41,** 703–707 (2009).
21. Jacob, C. O. *et al.* Lupus-associated causal mutation in neutrophil cytosolic factor 2 (NCF2) brings unique insights to the structure and function of NADPH oxidase. *Proc. Natl Acad. Sci. USA* **109,** E59–E67 (2012).
22. Kolaczkowska, E. & Kubes, P. Neutrophil recruitment and function in health and inflammation. *Nature Rev. Immunol.* **13,** 159–175 (2013).

23. Liu, D. J. & Leal, S. M. Estimating genetic effects and quantifying missing heritability explained by identified rare-variant associations. *Am. J. Hum. Genet.* **91,** 585–596 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements The study was primarily funded by the Medical Research Council (MRC G1001158 to D.A.v.H and V.P.), with further funding from Coeliac UK (to D.A.v.H). We thank C. Wijmenga and G. Trynka for sharing ImmunoChip data, and the International Multiple Sclerosis Genomics Consortium for ImmunoChip data and samples. J.N.B. and R.C.T. are supported by MRC grant G0601387. This research was supported by the National Institutes for Health Research (NIHR) Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. The study was supported by the Cambridge NIHR Biomedical Research Centre. We thank E. Gray and D. Jones (Wellcome Trust Sanger Institute) for sample preparation. We acknowledge use of DNA from The UK Blood Services collection of Common Controls (UKBS-CC collection), funded by the Wellcome Trust grant 076113/C/04/Z and by NIHR programme grant to NHS Blood and Transplant (RP-PG-0310-1002). The collection was established as part of the Wellcome Trust Case Control Consortium (WTCCC). We acknowledge use of DNA from the British 1958 Birth Cohort collection, funded by the UK MRC grant G0000934 and the Wellcome Trust grant 068545/Z/02. We thank nurses and doctors for recruiting autoimmune thyroid disease (AITD) subjects into the AITD National Collection, funded by the Wellcome Trust grant 068181. We acknowledge use of DNA from the Cambridge BioResource. We acknowledge use of DNA from the Juvenile Diabetes Research Foundation (JDRF)/Wellcome Trust Case-Series (GRID), funded by JDRF and the Wellcome Trust (grant references JDRF 4-2001-1008 and WT061858). The subjects were recruited in the UK by D. Dunger and his team with support from the British Society for Paediatric Endocrinology and Diabetes. The samples were prepared and provided by the JDRF/Wellcome Trust Diabetes and Inflammation Laboratory, University of Cambridge, UK. Psoriasis samples used were based on the WTCCC2 GWAS clinical panel, for which we thank D. Burden, C. Griffiths, M. Cork and R. McManus. Finally, we would like to thank all autoimmune disease and control subjects for participating in this study.

Author Contributions D.A.v.H. designed and led the study. K.A.H. coordinated wet laboratory work, with K.A.H., V.M., N.A.B. and E.W. performing DNA sample preparation, Fluidigm PCR amplification, sample barcoding, MiSeq library validation and Sanger sequencing preparation. HiSeq sequencing was performed by M.M. and E.P. D.A.v.H., V.P. and M.S. performed bioinformatics and statistical analyses. All other authors contributed to diverse aspects of sample collection, phenotyping, DNA preparation, ImmunoChip data production or specific analyses. D.A.v.H. and V.P drafted the manuscript, which all authors reviewed.

Author Information Genome data has been deposited at the European Genome–phenome Archive (http://www.ebi.ac.uk/ega/), which is hosted at the EBI, under accession number EGAS00001000476. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.A.v.H. (d.vanheel@qmul.ac.uk) or V.P. (v.plagnol@ucl.ac.uk).

METHODS

Gene selection. All genes studied (listed in Supplementary Table 3) were risk loci for at least two phenotypes, had a known immune system function, were from loci with only a single strong candidate immune gene (or all immune genes were selected at four loci: *IL18R1*, *IL18RAP*; *CTLA4*, *CD28*, *ICOS*; *IL2*, *IL21*; *PTPRK*, *THEMIS*), and all genes and loci were densely genotyped with all 1000 Genomes pilot project variants on the Illumina ImmunoChip (for design of this chip, see ref. 19). Additional criteria favouring locus selection were: known multiple independent association signals, risk (not necessarily same variants/haplotype or signal direction) for many autoimmune diseases, fine-mapping or other data strongly suggesting a single candidate gene, and smaller complementary DNA size.

Samples. UK samples for the six component immune disease phenotypes have been described in previous publications (which also contain full details of Ethics Committee approvals)[19,20,24-27], as have the three control populations[19,28]. Informed consent was obtained from all subjects. Individuals with self-reported autoimmune disease were excluded from the UK Blood Services — Common Controls and NIHR Cambridge Biomedical Research Centre Cambridge BioResource controls. Samples with self-stated non-white European ethnicity were excluded (later further confirmed by ImmunoChip-based principal component ethnicity analysis for 32,806 samples). Samples with gross discordance with ImmunoChip genotypes and/or with known gender or genotype-mismatch issues from previous GWAS were excluded. Samples with known duplicates or relatedness (as distant as first cousins) were excluded, relatedness was later confirmed by ImmunoChip genome-wide identity-by-state analysis and by analysis of multiple rare-variant sharing in Fluidigm sequencing data. Additional independent European samples genotyped for rs17849502 (4,313 coeliac cases and 3,954 controls) were previously described[19].

Wet-lab. PCR primers were designed for all RefSeq exons of 26 genes, and amplicons selected to be 150–200 base pairs (bp) in size. There was minor primer design dropout at *IL18R1*, *STAT4*, *THEMIS* and *ZMIZ1*, although >94% of exon sequence was still covered at these genes. Variant calls at the gene *YDJC* later proved unreliable with highly biased allele depths at heterozygote sites, probably due to the very high exon GC content (~70%), and this gene was not further analysed nor is it discussed elsewhere in this study. The total length of (overlapping) amplicons was 95,927 bp; with primers removed (still overlapping) 72,612 bp; and with primers removed and unique sequence 58,550 bp. PCR amplification was performed using 50 ng genomic DNA per sample on the 48 sample/plate Fluidigm microfluidic Access Array system. PCR primers for 511 PCR reactions were pooled up to 12-plex per well in 48 pools. Individual per sample per pool PCR reactions took place in ~35-nl reaction chambers with ~300 DNA haplocopies per reaction. All pools per sample were combined. Each sample's pool was then individually barcoded in a second PCR reaction with one of 1,536 10-bp Fluidigm-designed unique barcodes (Fluidigm unidirectional sequencing protocol).

Sequencing. Thirty-four libraries (each of 1,536 barcoded samples) were generated. Libraries were first sequenced on an Illumina MiSeq for rapid quality control of the barcoding step, and to optimize loading concentrations/cluster density. Libraries were then sequenced one per lane using 101-bp paired-end reads and an 11-bp index read (the last base of each read being only used for chemistry cycle phasing purposes) on Illumina HiSeq sequencers. Lanes were repeated if target cluster density or target clusters passing filter were not achieved. Individual samples were de-multiplexed by Illumina CASAVA software, allowing zero mismatches per 10-bp barcode. Sanger sequencing was performed on PCR products using an ABI 3730xl DNA analyser and ABI big dye terminator 3.1 cycle chemistry. We sequenced all samples with rare-variant allele genotypes, and a control sample, for the 74 sites selected.

Bioinformatics. PCR primers were trimmed from the 5′ end of individual reads using a modified version of btrim[29]. Trimmed sequences were aligned to the GRCh37 human reference genome using gapped quality-aware alignment, and base call quality recalibration implemented in Novoalign V2.07.18 with settings '-t 100 -H -g 65 -x 7 -o FullNW'. Data were realigned against known (1000 Genomes and Mills-Devine 2-hit) indels and per-sample called indels. SNPs were called using GATK 1.6-5 and settings '–min_base_quality_score 15 -stand_call_conf 30–baq CALCULATE_AS_NECESSARY -glm SNP–baqGapOpenPenalty 65–downsampling_type BY_SAMPLE–downsample_to_coverage 250' and then hard filtered using GATK settings 'QUAL<80.0 DP<20 MQ<40.0 QD<2.0 MQRankSum<-12.5 HRun>5′ (several other recommended best practice GATK settings were not appropriate for PCR amplicon data), and around indels. Small indels (up to 15-bp gaps from Novoalign) were called using GATK and settings '–min_base_quality_score 15 -stand_call_conf 30–baq CALCULATE_AS_NECESSARY -glm INDEL–baqGapOpenPenalty 65–downsampling_type BY_SAMPLE–downsample_to_coverage 250' and then hard filtered using GATK settings 'QUAL<80.0 DP<20 QD<2.0' (several other recommended best-practice GATK settings were not appropriate for PCR amplicon data). The most important of these settings were likely to be calling genotypes as missing with sequencing depth <20 high-quality bases and the minimum Phred 15 recalibrated base call quality score to define high-quality bases. Both SAMtools and VCFtools software were also used to process data. SNP genotypes (including non-reference genotypes) were called at all 58,550 bases of amplicon sequence. Samples with <57,600 SNP genotype calls (98.4%, a threshold determined by inspection of the call rate plot) were removed and scheduled for repeat processing. Clusters of very close non-reference genotypes in an individual sample were removed. Non-reference genotype sites were then identified across all samples, and VCF-level data reduced to variants at polymorphic sites (in one or more samples). A combined VCF file of all polymorphic sites and samples was then loaded into PLINK/SEQ v0.09. Multiple-step filtering based on call rate per sample and call rate per variant site was applied, with final requirements >99.95% call rate per sample and per variant site. Lower call rate samples at this stage were also scheduled for repeat processing. We removed variants if the sum of heterozygote genotype allele depths was <25% or >75%. The final filtered data was then exported to a VCF file containing all variants and samples for analysis in R. ImmunoChip data was loaded into Illumina GenomeStudio software from .idat files, and all samples called together in GenomeStudio using the cluster settings as previously described[19]. Data were merged with HapMap Phase 3 genotypes, principal component analysis performed, and the first two principal components used to validate ethnicity (Supplementary Fig. 3).

Barcode and sequencing amplicon performance. Barcode evenness was excellent, with typically 99.0% of the 1,536 barcodes producing pass-filter read numbers that were between 0.033% and 0.13% of the total pass-filter reads per lane (0.065% expected), with most of the failing barcodes tagging known water-negative control samples or (based on repeat amplification with a different barcode) due to poor DNA quality. Amplicon evenness was good, and for many genotype calls we were required to downsample data to 250 bases per site per sample (Supplementary Fig. 1). However, 10 of 511 amplicons effectively failed PCR. In a typical analysis of 100 high-quality samples, 2% of the 58,550 unique amplicon bases had a minimum mean read-depth of <20, nearly all accounted for by the 10 failing amplicons.

Variant annotation. Annotation of all variants was first performed using ANNOVAR (Feb 2013) and the GENCODE V14 data set. Coding variants were identified. Rare functional variants were identified based on stop, frameshift indel, nonsynonymous (SNV or 3*n* indel) or splice predictions. We performed an additional layer of annotation for high confidence loss of function mutations, using the methods described in ref. 30. The Variant Effect Predictor (VEP v2.5) tool from Ensembl was modified to produce custom annotation tags and additional loss of function (LOF) annotations. The additional LOF annotation was applied to variants which were annotated as STOP_GAINED, SPLICE_DONOR_VARIANT, SPLICE_ACCEPTOR_VARIANT, and FRAME_SHIFT and flagged if any filters failed. Filters included: LOF is the ancestral allele; exon is surrounded by non-canonical splice site (that is not AG/GT); LOF removes less than 5% of remaining protein; LOF is rescued by nearby start codon which results in less than 5% of protein truncated; transcript only has one coding exon; splice-site mutation within intron smaller than 15 bp; splice site is non-canonical OR other splice site within same intron is non-canonical; unable to determine exon/intron boundaries surrounding variant. A LOF variant is predicted as high confidence if there is one transcript that passes all filters, otherwise it is predicted as low confidence. We noted that LOF mutations were seen in 21 out of 25 genes, all were heterozygous genotypes, and mainly (87 out of 97) as singletons or doubletons in the 41,911 samples (Supplementary Table 3).

Statistical analysis. Most analysis was performed in R using custom code (available on request). For tests using permutations (C-alpha, UNIQ-cases and UNIQ-controls in Fig. 1), we randomly permuted in R the case–control status 10,000 times. The unconditional burden test (Fig. 1b) used a Fisher's exact text. Conditional burden tests used the glm function in R, including selected ImmunoChip common variants as covariates (selection based on a stepwise regression analysis up to 10^{-4}). For the C-alpha statistic computation (Fig. 1a), the expected proportion of rare alleles in the case–control cohorts was set to the proportion of cases and controls. Figure 1 was generated using the fact that under the null of no association $-2\log(P)$ is distributed as chi-squared with 2 degrees of freedom. PLINK/SEQ v0.09 (http://atgu.mgh.harvard.edu/plinkseq/index.shtml) was used for Ti/Tv statistics, and to confirm findings of R analyses (not shown). We used PLINK/SEQ for the genotype concordance analysis between ImmunoChip and Fluidigm-sequencing data. Discordant calls were observed at 169 of 2,985,255 (0.0056%) genotypes, occurring at 36 out of 91 polymorphic variant sites present in both data sets. We inspected Illumina ImmunoChip R theta intensity plots for the discordant genotypes, and observed 8 discordant genotypes to be likely due to ImmunoChip data mis-clustering, and 11 discordant genotypes to be due to a third or fourth observed allele in the high-throughput sequencing data. At the sites with third and fourth alleles, we note the ImmunoChip array assays can

only call two alleles, therefore is not possible to determine whether these sequence genotype calls are real or errors. R code used for analysis is available from V.P.

Estimation of average genetic effect contributed by rare variants. For each combination of locus by disease, we combined all rare functional variants (frequency $< 0.5\%$ in 1,000 Genomes/NHLBI data sets and nonsynonymous, LOF or splicing) in a burden statistic X and computed the combined frequency of X in the sample. Using a logistic regression model with the disease phenotype as outcome, we estimated the odds ratio associated with the burden variable X. This knowledge of frequency and odds ratio for the burden variable X enables the estimation of the average genetic effect (AGE, as defined in ref. 23) version of the variance explained. We then compared this variance at each combination of locus/gene with the variance explained by what we consider to be a typical common variant association (odds ratio 1.2, MAF 20%, assuming a single common variant per locus). To deal with the uncertainty in estimated odds ratio and obtain a confidence interval for this value, we randomly sampled the odds ratio from their estimated distribution for each pair of disease/locus. Averaging over the 150 combinations of 6 diseases by 25 loci, we estimate the ratio of heritability explained for all rare variants by all common variants to have a mean value of 1.6%, with a confidence interval of (1.2–2.3%). It is pointed out in ref. 23 that the AGE estimate can underestimate the true explained variance by rare variants. Nevertheless, assuming that rare variants are generally all risk or all protective at a given gene, their simulations also show that the underestimation is limited, in the range of a 25% decrease. Taking this conservative estimate of the under-estimation level, we find the upper bound of the 95% of the confidence interval to be 3.05%. Hence, our data indicate that the aggregate contribution of rare variants to the heritability ($<0.5\%$ MAF, and averaged over these loci/diseases) is unlikely to exceed approximately 3% of the heritability assigned to common variants. We acknowledge that a much larger underestimation (and therefore a much larger heritability explained for rare variants) is possible in the presence of a combination of high risk and highly protective rare variants at the same locus. Although we cannot exclude such scenario, it is unlikely to be widespread. We also assumed in our estimates that rare variants act additively at the log scale. Although this assumption is standard, we cannot exclude that a combination of rare variants results in a much stronger predictive outcome than rare variants individually, hence underestimating the heritability associated with rare variants.

24. Cooper, J. D. *et al.* Seven newly identified loci for autoimmune thyroid disease. *Hum Mol Genet* **21,** 5202–5208 (2012).
25. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491,** 119–124 (2012).
26. Sawcer, S. *et al.* Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476,** 214–219 (2011).
27. Tsoi, L. C. *et al.* Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nat Genet* **44,** 1341–1348 (2012).
28. Dendrou, C. A. *et al.* Cell-specific protein phenotypes for the autoimmune locus IL2RA using a genotype-selectable human bioresource. *Nat Genet* **41,** 1011–1015 (2009).
29. Kong, Y. Btrim: a fast, lightweight adapter and quality trimming program for next-generation sequencing technologies. *Genomics* **98,** 152–153 (2011).
30. MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335,** 823–828 (2012).

Published in final edited form as:
Nat Genet. ; 43(12): 1193–1201. doi:10.1038/ng.998.

Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease

Gosia Trynka[1,35], **Karen A Hunt**[2,35], **Nicholas A Bockett**[2], **Jihane Romanos**[1], **Vanisha Mistry**[2], **Agata Szperl**[1], **Sjoerd F Bakker**[3], **Maria Teresa Bardella**[4,5], **Leena Bhaw-Rosun**[6], **Gemma Castillejo**[7], **Emilio G. de la Concha**[8], **Rodrigo Coutinho de Almeida**[1], **Kerith-Rae M Dias**[6], **Cleo C. van Diemen**[1], **Patrick CA Dubois**[2], **Richard H. Duerr**[9,10], **Sarah Edkins**[11], **Lude Franke**[1], **Karin Fransen**[1,12], **Javier Gutierrez**[1], **Graham AR Heap**[2], **Barbara Hrdlickova**[1], **Sarah Hunt**[11], **Leticia Plaza Izurieta**[13], **Valentina Izzo**[14], **Leo AB Joosten**[15,16], **Cordelia Langford**[11], **Maria Cristina Mazzilli**[17], **Charles A Mein**[6], **Vandana Midah**[18], **Mitja Mitrovic**[1,19], **Barbara Mora**[17], **Marinita Morelli**[14], **Sarah Nutland**[20], **Concepción Núñez**[8], **Suna Onengut-Gumuscu**[21], **Kerra Pearce**[22], **Mathieu Platteel**[1], **Isabel Polanco**[23], **Simon Potter**[11], **Carmen Ribes-Koninckx**[24], **Isis Ricaño-Ponce**[1], **Stephen S. Rich**[21], **Anna Rybak**[25], **José Luis Santiago**[8], **Sabyasachi Senapati**[26], **Ajit Sood**[18], **Hania Szajewska**[27], **Riccardo Troncone**[28], **Jezabel Varadé**[8], **Chris Wallace**[20], **Victorien M Wolters**[29], **Alexandra Zhernakova**[30], **CEGEC (Spanish Consortium on the Genetics of Coeliac Disease)**, **PreventCD Study Group**, **Wellcome Trust Case Control Consortium**, **B.K. Thelma**[26], **Bozena Cukrowska**[31], **Elena Urcelay**[8], **Jose Ramon Bilbao**[13], **M Luisa Mearin**[32], **Donatella Barisani**[33], **Jeffrey C Barrett**[11], **Vincent Plagnol**[34], **Panos Deloukas**[11], **Cisca Wijmenga**[1,36], and **David A van Heel**[2,36]

[1]Genetics Department, University Medical Center and University of Groningen, PO Box 30.001, 9700 RB Groningen, The Netherlands [2]Blizard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London E1 2AT, United Kingdom [3]Department of Gastroenterology, VU Medical Center, 1007 MB Amsterdam, The Netherlands [4]Fondazione IRCCS Ospedale Maggiore Policlinico, Mangiagalli e Regina Elena, Milan, Italy. [5]Department of Medical Sciences, University of Milan, Milan, Italy. [6]Genome Centre, Barts and the London School of Medicine and Dentistry, John Vane Science Centre, Charterhouse Square, London, EC1M 6BQ, United Kingdom [7]Universitat Rovira I Virgili, Department of Paediatric Gastroenterology, Hospital Univesitari de Sant Joan de Reus, , 43201 Reus, Spain [8]Immunology Dept, Hospital Clínico S. Carlos, Instituto de Investigación Sanitaria San Carlos IdISSC, Madrid,

Correspondence to DAvH (d.vanheel@qmul.ac.uk) and CW (c.wijmenga@medgen.umcg.nl).
[35]These authors contributed equally to this work
[36]These authors jointly directed this project.

AUTHOR CONTRIBUTIONS DAvH and C. Wijmenga led the study. Major contributions were (i) DAvH, KAH, GT and C. Wijmenga wrote the paper; (ii) KAH, GT, VM, NB, JR, MP, MM, RHD and KF performed DNA sample preparation and genotyping assays; (iii) DAvH, VP, KAH, GT performed statistical analysis. Other authors contributed mainly to sample collection and phenotyping. PD led the formation of the Immunochip Consortium, with SNP selection by JB and C. Wallace. All authors reviewed the final manuscript.

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.

URLs
Database of Genomic Variants, http://projects.tcag.ca/variation/?source=hg18
T1Dbase: www.t1dbase.org
SIFT: sift.jcvi.org
BioGPS: biogps.gnf.org

URLs for Consortia and Groups
www.preventcd.com
www.wtccc.org.uk

Europe PMC Funders Author Manuscripts

Spain [9]Division of Gastroenterology, Hepatology and Nutrition, Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA [10]Department of Human Genetics, University of Pittsburgh Graduate School of Public Health, Pittsburgh, Pennsylvania, USA [11]Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, United Kingdom [12]Department of Gastroenterology, University Medical Center and Groningen University, 9700 RB Groningen, The Netherlands [13]Immunogenetics Research Laboratory, Hospital de Cruces, Barakaldo 48903 Bizkaia, Spain [14]European Laboratory for Food Induced Disease, University of Naples Federico II, Naples, Italy. [15]Department of Medicine, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands [16]Nijmegen Institute for Infection, Inflammation and Immunity (N4i), Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands [17]Department of Molecular Medicine, Sapienza University of Rome, Rome, Italy [18]Dayanand Medical College and Hospital, Ludhiana, Punjab, India [19]University of Maribor, Faculty of Medicine, Center for Human Molecular Genetics and Pharmacogenomics, Slomskov trg 15, 2000 Maribor, Slovenia [20]Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory, Department of Medical Genetics, Cambridge Institute for Medical Research, University of Cambridge, Cambridge CB2 0XY, United Kingdom [21]Center for Public Health Genomics, University of Virginia, Charlottesville, VA 22908-0717 [22]UCL Genomics, Institute of Child Health, University College London, 30 Guilford Street, London WC1N 1EH, United Kingdom [23]Pediatrics Gastroenterology Department, Hospital La Paz, Madrid, Spain [24]La Fe University Hospital, Pediatric Gastroenterology, Bulevar Sur s/n 46026 Valencia, Spain [25]Department of Gastroenterology, Hepatology and Immunology, Children's Memorial Health Institute, Warsaw, Poland [26]Department of Genetics, University of Delhi, South Campus, New Delhi, India. [27]The Medical University of Warsaw, Department of Pediatrics, Dzialdowska 1, 01-184 Warsaw, Poland [28]University of Naples, Fedrico II, Department of Pediatrics, Via S.Pansini 5, 80131 Naples, Italy [29]Department of Paediatric Gastroenterology, University Medical Centre Utrecht, Utrecht, The Netherlands [30]Department of Rheumatology, Leiden University Medical Center, Leiden, The Netherlands [31]Department of Pathology, Children's Memorial Health Institute, Warsaw, Poland [32]Department of Paediatrics, Leiden University Medical Centre, Leiden, The Netherlands [33]Department of Experimental Medicine, Faculty of Medicine University of Milano-Bicocca,Monza, Italy [34]UCL Genetics Institute, University College London, Gower Street, London WC1E 6BT

Abstract

We densely genotyped, using 1000 Genomes Project pilot CEU and additional re-sequencing study variants, 183 reported immune-mediated disease non-*HLA* risk loci in 12,041 celiac disease cases and 12,228 controls. We identified 13 new celiac disease risk loci at genome wide significance, bringing the total number of known loci (including *HLA*) to 40. Multiple independent association signals are found at over a third of these loci, attributable to a combination of common, low frequency, and rare genetic variants. In comparison with previously available data such as HapMap3, our dense genotyping in a large sample size provided increased resolution of the pattern of linkage disequilibrium, and suggested localization of many signals to finer scale regions. In particular, 29 of 54 fine-mapped signals appeared localized to specific single genes - and in some instances to gene regulatory elements. We define a complex genetic architecture of risk regions, and refine risk signals, providing a next step towards elucidating causal disease mechanisms.

INTRODUCTION

Celiac disease is a common complex chronic immune-mediated disease with seroprevalence of ~1%[1,2] in individuals of white European origin. A T-cell mediated small intestinal

immune response is generated against gliadin fragments from wheat, rye and barley cereal proteins leading to villous atrophy. Its aetiology is poorly understood. Association with *HLA* variants was first shown in 1972, and predisposing *HLA-DQ2* and *-DQ8* sub-types are necessary but not sufficient to cause disease. Recent genome wide association studies (GWAS) have identified a further 26 non-*HLA* risk loci[3-6]. Many of these loci are also associated with other autoimmune or chronic immune-mediated diseases (albeit sometimes different markers and effect directions[7]), with particular overlap observed between celiac disease, type 1 diabetes[8] and rheumatoid arthritis[9].

Currently unanswered questions regarding the genetic predisposition to celiac disease, which are also relevant for other immune-mediated diseases, include explaining the remaining major fraction of heritability, including rare and additional common risk variants; and identification of causal variants and causal genes (or at least more finely localizing the risk signal). The Immunochip Consortium[10] developed to explore these questions, taking advantage of emerging comprehensive common, low frequency, and rare variation datasets, and of a commercial offer of much lower per-sample custom genotyping costs for a very large project comprised of related diseases.

The Immunochip, a custom Illumina Infinium HD array, was designed to densely genotype, using 1000Genomes and any other available disease specific resequencing data, immune-mediated disease loci identified by common variant GWAS. The 1000 Genomes Project pilot CEU low-coverage whole genome sequencing dataset captures 95% of variants of MAF=0.05, and although underpowered to comprehensively detect variants of rarer allele frequency, still identifies 60% of variants of MAF=0.02, and 30% of variants of MAF=0.01[11]. The Consortium selected 186 distinct loci containing markers meeting genome wide significance criteria ($P<5\times10^{-8}$) from twelve such diseases (autoimmune thyroid disease, ankylosing spondylitis, Crohn's disease, celiac disease, IgA deficiency, multiple sclerosis, primary biliary cirrhosis, psoriasis, rheumatoid arthritis, systemic lupus erythematosus, type 1 diabetes and ulcerative colitis). All 1000 Genomes Project low-coverage pilot CEU population sample variants[11] (Sept 2009 release) within 0.1cM (HapMap3 CEU) recombination blocks around each GWAS region lead marker were submitted for array design. No filtering on correlated variants (linkage disequilibrium) was applied. Further case and control regional resequencing data were submitted by several groups (Online Methods, Supplementary Note), as well as a small proportion of investigator-specific undisclosed content including intermediate-significance GWAS results.

Most GWAS have been performed using common SNPs (typical minor allele frequency (MAF) >5%), further selected for low inter-marker correlation and/or even genomic spacing. In contrast to GWAS, the Immunochip presents a comprehensive in-depth opportunity to dissect the architecture of both rare and common genetic variation, at immuno-biologically relevant genomic regions, in human diseases. Due to the presence in our final Immunochip dataset of the majority of 1000 Genomes Project pilot CEU polymorphic genetic variants (and additional resequencing at some loci), the true causal variants from many risk loci may have been directly genotyped and analysed.

RESULTS

A total of 207,728 variants were submitted for Immunochip assay design and 196,524 passed manufacturing quality control at Illumina. After extensive and stringent data quality control (Online Methods), we analysed a near-complete dataset (overall 0.008% missing genotype calls) comprising 12,041 celiac disease cases and 12,228 controls (from 7 geographic regions, Table 1) and 139,553 polymorphic (defined here as ≥2 observed genotype groups) markers. 634 biallelic SNPs were assayed in duplicate, at these we

observed 189 of 15,384,884 (0.0012%) genotype calls to be discordant. Considering the intended 207,728 variants submitted for design, and an observed ~9.1% non-polymorphic rate in our post-quality control data, we estimate we have high quality genotype data on ~74% of the complete 1000 Genomes Project pilot CEU true polymorphic variant set at the fine-mapped regions.

We observed that 36 of the 183 non-HLA immune-mediated disease loci selected for Immunochip dense 1000Genomes-based genotyping achieved genome-wide significance ($P<5\times10^{-8}$) for celiac disease in either the current study or our previous GWAS[5] (summary association statistics for all markers are available in T1DBase). All variants reaching genome wide significance were common (MAF >5%). We also observed marked enrichment for intermediate significance level celiac disease association signals (e.g. rs6691768, *NFIA* locus, P= 5.3×10^{-8}) at a proportion of the remaining 147 dense-genotyped non-celiac autoimmune disease regions (Supplementary Figure 1). Variants from 3 dense-genotyped regions selected on Immunochip for a non-immune-mediated trait (bipolar disorder) showed no excess of association signals (Supplementary Figure 1).

We identified 13 new celiac risk loci ($P<5\times10^{-8}$, Figure 1, Table 2, Supplementary Figure 2), 10 of which were from immune-mediated disease loci selected for Immunochip dense 1000Genomes-based genotyping. Several of these new loci were reported at lesser significance levels in our previous studies[5,9], and almost all have been reported in at least one other immune-mediated disease. These, with *HLA*, bring to 40 the total number of reported (current and/or previous study[5], which had an overlapping but slightly different sample set) genome wide significant celiac disease loci. Most contain candidate genes of immunological function, consistent with our previous findings at celiac disease loci[3-5].

Effect sizes (odds ratios, inverting protective effects) for the most significant marker per locus were median 1.155 (range 1.124 – 1.360) for the top signals from 26 non-HLA loci measured using Illumina Hap300/Hap550-chip linkage disequilibrium-pruned tag SNPs in our 2010 celiac disease GWAS[5] and median 1.166 (range 1.087 – 1.408) for the corresponding most significant marker (for the same signal) per locus in the current high density fine-mapping Immunochip dataset (Wilcoxon test $P=0.75$, Supplementary Table 1). Although we observe no difference in effect sizes between GWAS lead SNPs and subsequent fine-mapped signals, we note that case resequencing in the current Immunochip dataset is limited (see also **Discussion**).

In all, we report 57 independent coeliac disease association signals (Table 2) from 39 separate loci, of which 18 (32%) were not efficiently ($r^2>0.9$, Supplementary Table 2) tagged by our previous GWAS[5] (Illumina Hap550, post quality control dataset) markers.

Multiple independent common and rare variant signals

In contrast to most GWAS chips, the Immunochip contains a substantial proportion of lower MAF polymorphic variants. Of 139,553 variants in our 11,837 European-origin controls, 24,661 variants are low frequency (defined[11] as MAF 5% to 0.5%) and a further 22,941variants are rare (MAF<0.5%). We investigated the possibility of multiple independently associated variants (of all allele frequencies) at each locus, using stepwise logistic regression conditioning on the most significant variant at the locus (Online Methods, Supplementary Table 3). This analysis can be sensitive to genotype miscalling and missing data[12], hence our use of extremely rigorous quality control measures for the dataset and manual inspection of genotype clusters for all reported markers.

We observed two or more independent signals at 13 of 36 high-density genotyped non-HLA loci (Figure 2). Four of these loci each had three independent signals (*STAT4*, the

chromosome 3 *CCR* region, *IL12A*, *SOCS1*/*PRM1*/*PRM2*, Table 2). Low frequency and/or rare variant signals were seen at four separate loci (*RGS1*, *CD28*/*CTLA4*/*ICOS*, *SOCS1*/*PRM1*/*PRM2*, *PTPN2*). Notably, the strongest effect (OR 1.70) was seen at the rare variant imm_16_11281298 (*SOCS1*/*PRM1*/*PRM2* locus) with genotype counts (AA/AG/GG) of 1/136/11904 (MAF 0.57%) in all celiac cases and 0/91/12136 (MAF 0.37%) in all controls (detailed genotype count and allele frequency data for top signals by collection are shown in Supplementary Table 4).

We next performed haplotype analysis on all loci with multiple independent signals, to investigate whether the multiple signals were due to multiple causal effects or a single effect best tagged by several variants. For all but one locus (*PTPN2*) the haplotype association tests (not shown) were of similar significance to the single SNP association tests, suggesting that for each signal we have genotyped either the causal variant, or markers very strongly correlated with it. These findings contrast with those from a recent resequencing study[13], probably because of the much greater variant density of our study. However, at the *PTPN2* locus, the imm_18_12833137(T) + ccc-18-12847758-G-A(G) haplotype was considerably more associated (P=4.8×10^{-14}, OR 0.84) than either SNP alone (imm_18_12833137 P=1.9×10^{-10}; ccc-18-12847758-G-A P= 0.0008).

Interestingly at the *SOCS1* locus, the third independent signal imm_16_11292457 shows association only after conditioning on the two other signals (P=2.0×10^{-4}) but not in the single SNP non-conditioned association analysis (P=0.15). Further inspection revealed the protective imm_16_11292457(A) allele to be correlated (in linkage disequilibrium) with the risk (A) allele of the first signal imm_16_11268703, thus although there are indeed three independent signals, the effect of the third signal is only revealed after conditioning on the first. A similar statistical effect (Simpson's paradox) was recently shown at a Parkinson's disease locus[14].

Fine-mapping to localize causal signals

GWAS signals are typically reported within relatively large linkage disequilibrium blocks. We tested whether our much denser genotyping strategy would allow finer-scale localization, and the pinpointing of association signals. We found that markers strongly correlated (r^2>0.9) with the most significant independent variant clustered together, and defined regions that are a median 12.5x smaller than the relevant HapMap3 CEU 0.1cM linkage disequilibrium blocks (Table 2, Figure 2, Supplementary Figure 2). Localization was highly successful for some regions (e.g. *PTPRK, TAGAP*), but not possible at others (e.g. *IL2-IL21*). At many loci, the localized regions comprised only a handful of markers in close physical proximity.

Considering the 36 high density genotyped loci, we have localized to a single gene 29 of the total 54 independent non-*HLA* signals (Table 2, Supplementary Figure 2). We identified all markers strongly correlated (r^2>0.9) with the independent non-*HLA* variants reported in our analyses (from Table 2), and on functional annotation (Supplementary Table 2) identified only a handful of markers in exonic regions and of these only three are protein altering variants (nsSNPs: imm_1_2516606 (*MMEL1*), imm_12_110368991 (*SH2B3*), 1kg_X_152937386 (*IRAK1*). In contrast, a number of signals appeared to be more finely localized around the transcription start site of specific genes (which we defined as the first exon, and 10kb 5′ of the first exon), including signals at *RUNX3*, *RGS1*, *ETS1*, *TAGAP*, *ZFP36L1*; and around the 3′ UTR region (and 10kb 3′) including signals at *IRF4*, *PTPRK* and *ICOSLG*.

Overlap between multiple independent signal regions was seen at some loci (Figure 2), suggesting that causal variants might be functioning through a shared mechanism e.g. within

a 2kb region of the *PTPRK* 3′ UTR; within a 11kb region 5′ of *IL12A*; or within a 28kb region of *TNFAIP3*. In contrast, multiple independent signals were observed that spread between the three immune genes of the *CD28/CTLA4/ICOS* region.

DISCUSSION

We show that fine mapping of GWAS regions using dense resequencing data, e.g. (as here) from the 1000Genomes project, is feasible and generates substantial additional information at many loci. We identify a complex architecture of multiple common and rare genetic risk variants at around a third of the now 40 proven celiac disease loci. The design of our study has allowed us to find many more such complex regions than the ~10% with multiple signals seen in our previous study[5] and a recent large GWAS for human height[15]. It seems probable that if larger sample sizes than in the current study were to be tested, additional loci might be shown to have a similarly rich multiple risk variant architecture. Multiple independent risk signals for celiac disease have also long been known in the *HLA* region[16]. Our success in celiac disease might be partly due to the extensive selective pressures for haplotypic diversity that have taken place at immune gene loci[17]. Previous studies have reported independently associated common and rare variants at individual loci for a handful of phenotypes e.g. fetal haemoglobin[13], sick sinus syndrome[18], Crohn's disease[19], hypertriglyceridemia[20]. To the best of our knowledge, ours is the first study to have comprehensively surveyed the genetic architecture of all known risk loci for a trait.

In part, our identification of rare variants at risk regions relies on the prior discovery of a genome-wide significant common variant association signal at each locus. This then permits a per-locus rather than genome-wide multiple testing correction when searching for additional independent association signals. Only particularly strong rare variant signals would, on their own, generate significance levels reaching the genome-wide threshold typically used in GWAS studies ($P<5\times10^{-8}$). Alternative methods, such as collapsing rare variant signals across a gene or functional categories of genes have therefore been suggested as approaches to the same problem[21]. Although a rare variant may have occurred on a recent haplotypic background, and thus show linkage disequilibrium at substantially longer range than common variants, we deliberately restricted our search to around the common variant linkage disequilibrium blocks as to do otherwise would have incurred a considerably greater penalty from multiple testing. Therefore, although our study provides considerable encouragement for exome and whole genome sequencing efforts aimed at identifying rare risk variants (not necessarily restricted to GWAS loci) in common complex diseases, it further highlights the statistical challenges of establishing rare variant associations.

We used a dense genotyping strategy and stepwise conditional association analysis, but did not identify any rare highly penetrant variants that might explain the genome-wide significant common SNP signals at any of the 39 loci. Our study does have limitations in this regard, particularly i) analysis restricted to 0.1cM linkage disequilibrium blocks; ii) the limited control resequencing sample size of the 1000 Genomes Project pilot CEU dataset; iii) the limited case resequencing sample size; and iv) case resequencing limited to three loci for celiac disease, and selected loci for other immune diseases. We observed a weak trend towards lower MAF ($P=0.042$, Wilcoxon test, Supplementary Table 1) for the best fine-mapping SNP (Immunochip experiment) versus the lead SNP from our 2010 tag SNP GWAS (measuring MAF in a subset of samples genotyped in both datasets). One signal showed substantially higher MAF (>25% change) on fine-mapping, four signals showed substantially lower MAF on fine mapping (Supplementary Table 1), yet all fine-mapping variants corresponding to lead GWAS SNPs remained common (MAF>0.10). We suggest that these changes in MAF upon fine-mapping of lead GWAS SNPs simply reflect more precise measurement of common frequency risk haplotypes. Although we cannot exclude

Europe PMC Funders Author Manuscripts

the possibility that a single high-penetrance lower-frequency variant explains most of the association signal at a locus, especially without more comprehensive case resequencing, we find no evidence in support this possibility in the current fine-mapping experiment. Nor can our stepwise selection procedure robustly refute the "synthetic association" hypothesis - in particular that a combination of multiple rare variants jointly explains the association signal[22] - although similarly we have not observed so far evidence supporting this possibility.

We established at genome wide significance 13 new loci for celiac disease, most of which have been reported previously at lesser significance or for another immune-mediated disease. The Illumina Hap550 chip (used in our 2010 GWAS) should have detected 10 of the 13 new loci, and in total 39 of the 57 independent non-HLA signals that we report. A current genotyping platform, the Illumina Omni2.5 chip would have detected 12 of the 13 new loci, and in total 50 of the 57 independent non-HLA signals that we report. Neither chip would have provided the finer scale localization of the Immunochip. The thirteen new loci contain many candidate genes of immunological function (P=0.0002 for enrichment of the Gene Ontology term "immune system process"[23]), in line with expectations from our previous studies. We also show evidence suggesting substantial additional signals at other immune-mediated disease loci, which lie beneath the genome wide significance reporting threshold applied to the current dataset. It is a point of debate whether such strict ($P<5\times10^{-8}$) criteria should apply - a Bayesian analyst might apply a higher prior at a locus already reported in another immune-mediated disease. Alternatively, an Immunochip-wide P value with a Bonferroni correction for independent SNPs, as used recently for the Cardiochip custom genotyping project[24], of $P<1.9\times10^{-6}$ (Online Methods) would yield 16 additional celiac disease loci. These 16 loci also mostly contain immune system genes. An analysis of these currently intermediate significance signals would gain substantial additional power by a meta-analysis across the several hundred thousand samples from multiple immune-mediated disease collections presently being run on Immunochip,

We found that our previous GWAS using tag SNPs gave very similar estimates of effect size to our current fine-mapping experiment (Supplementary Table 1), in contrast to a simulation study which suggested that GWAS markers often underestimate risk[14]. We have, however, found substantial evidence for multiple additional signals at known loci and report many new loci. In Europeans, the current 39 non-*HLA* loci now explain 13.7% of coeliac disease genetic variance (*HLA* accounts for a further ~40%). We also show a long tail of likely effects of weaker significance, which will explain substantial additional heritability.

Only one of the variants reported here was discovered by a disease-specific resequencing study: ccc-18-12847758-G-A (rs62097857), a marker identified by the WTCCC group's resequencing of Crohn's disease cases and controls (Supplementary Note) and also present in the Watson genome. We submitted for Immunochip ~4,000 variants from high throughput resequencing of pools of 80 celiac disease cases for extended genomic regions at three loci (*RGS1*, *IL12A*, *IL2- IL21*, Supplementary Note). These did not contribute additional signals over and above those obtained from the 1000 Genomes Project pilot CEU variants, although did contribute to increase the numbers of variants correlated with each signal (i.e the set of markers that likely contains the causal variant(s)) and more precisely define the bounds of the signal localization. We note that larger scale case resequencing (e.g. many hundreds of samples) would identify a rarer spectrum of variants than the current study, and has previously been used with success at selected genes and phenotypes.

The possibility of performing fine-scale mapping of GWAS regions using e.g. 1000 Genomes Project data has been discussed as a natural follow-on strategy for such studies[25,26] and has been recently used to identify risk variants in *APOL1* in African-

Americans with renal disease[27]. Our current report is the first to test such a strategy on a large scale in a complex disease. At multiple regions, we were able to refine the signal to a handful of variants over a few kilobases or tens of kilobases, although some regions (e.g. *IL2-IL21*) were resistant to this approach presumably due to particularly strong linkage disequilibrium. Most GWAS publications report signals mapping to a "LD block" based on HapMap recombination rates (sample size, 60 CEU families). In our data, where we have both i) much denser genotyping than GWAS chips (mean 13.6x at celiac loci versus the Illumina Hap550 chip) and ii) nearly 25,000 genotyped samples for the linkage disequilibrium calculations, we are able to observe much finer scale recombination and more precisely estimate of the bounds of no/minimal recombination intervals. Our findings are similar in terms of genotyping density and the resulting fine-mapped region size and lack of haplotype-specific effects to an earlier study of the *IL2RA* locus in type 1 diabetes[26]. At the majority of regions a tight block of highly correlated variants was seen, rather than a gradual decay of correlation (e.g. Figure 2 plots for *IL12A*, *PTPRK*). At many loci, we have now defined a handful of likely candidates to be the causal variant(s) to be taken forward into functional studies, although we may have missed candidate variants at some regions due to the sample size of the 1000 Genomes Project pilot CEU dataset (60 individuals), their status as controls, and our estimate that ~25% of these variants were excluded from our final dataset. These might be assessed by imputation methods[28], but our approach – particularly with regards to the more sensitive conditional regression analysis - has been to prefer the more accurate direct genotyping of all assayable variants. As and when much larger whole genome resequencing based reference datasets become available (e.g. the main 1000 Genomes Project), these might be used to impute into our Immunochip dataset, including substantially lower frequency variants[29]. We also investigated whether our use of multiple ethnic subgroups within Europe (e.g. southern European Spanish versus northern European UK) or the relatively small Indian collection contributed to fine mapping, and found that in most cases, the same degree of localization was possible with just the UK collection alone (data not shown).

Our data suggest that most common risk variants might function by influencing regulatory regions, consistent with those previously reported in other immune-mediated diseases, and complex traits in general[11]. The exception is the *SH2B3* nsSNP imm_12_110368991 (rs3184504), reported in our 2008 celiac GWAS[4], which even with the fine-mapping of 938 polymorphic variants from the *SH2B3* region remains the strongest signal at this locus thus suggesting it may be the causal variant. The same variant has been associated with other immune diseases, and a functional immune phenotype[5]. Interestingly, we observed a common ~980bp intergenic deletion between *IL2* and *IL21* (DGV40686, accurately genotyped by Infinium assay with control MAF 7.3%) correlated with the second independent signal at this region, although we have no evidence to suggest causality.

Our fine-scale localization approach has identified likely causal genes at many loci, and at eight genes signals localized around the 5′ or 3′ regulatory regions. For example, at the *THEMIS*/*PTPRK* locus, two independently associated sets of variants cluster in the 3′ UTR of the *PTPRK* gene (one, imm_6_128332892/rs3190930 in a predicted binding site for miRNA hsa-miR-1910). *PTPRK*, a TGF-beta target gene, is involved in $CD4^+$ T cell development and a deletion mutation causes T helper deficiency in the LEC rat strain[30]. The signal at *TAGAP* lies within a 4kb region immediately 5′ of the transcription start site, presumably containing promoter elements. At *ETS1*, the signal comprises 6 variants overlapping the promoter and 1st exon of the T cell expressed isoform NM_001162422.1, and one of the variants (imm_11_127897147/rs61907765) has predicted regulatory potential and overlaps multiple transcription factor binding sites (UCSC GenomeBrowser ChipSeq and ESPERR tracks, Supplementary Table 2). Similarly interesting variants are observed in regulatory regions of *RUNX3* (imm_1_25165788/rs11249212), and *RGS1*

(imm_1_190807644/rs1313292, imm_1_190811418/rs2984920) (Supplementary Table 2). Such an approach to identify the functional potential of risk variants was recently successful used to define a causal systemic lupus erythematosus *TNFAIP3* variant[31]. Although we have localized signals at many loci, and recent research suggests the likely causal gene is often located near the most strongly associated variant[15], only more detailed functional studies (e.g. transcription factor binding assays[31] and transcriptional activity assays of constructs with individual single nucleotide alterations at risk SNPs[32]), will prove precisely which gene variants might be causal.

We conclude that dense fine mapping of regions identified through GWAS studies can uncover a complex genetic architecture of independent common and rare variants, and often successfully localize risk variant signals to a small set of SNPs to be taken forward into functional assays. Denser fine mapping studies, utilising larger resequencing sample sizes from both cases and controls over broader regions, might provide further resolution of GWAS signals.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Coeliac UK for assistance with direct recruitment of celiac disease individuals, and UK clinicians (L.C. Dinesen, G.K.T. Holmes, P.D. Howdle, J.R.F. Walters, D.S. Sanders, J. Swift, R. Crimmins, P. Kumar, D.P. Jewell, S.P.L. Travis, K. Moriarty) who recruited celiac disease blood samples described in our previous studies. We thank the Dutch clinicians for recruiting celiac disease blood samples described in our previous studies (C.J. Mulder, G.J. Tack, W.H.M. Verbeek, R.H.J. Houwen, J.J. Schweizer). We thank the genotyping facility of the UMCG (Pieter van der Vlies) for helping in generating part of immunochip data and S. Jankipersadsing, A. Maatman, at UMCG for preparation of samples. We thank R. Scott for preparing samples for genotyping and the University of Pittsburgh Genomics and Proteomics Core Laboratories for performing genotyping. We thank C. Wallace for assistance with Immunochip SNP selection, and J. Stone for co-ordinating Immunochip design and production at Illumina. We thank the members of each disease consortium who initiated and sustained the cross-disease Immunochip project. We especially thank all individuals with celiac disease and control individuals for participating in this study.

Funding was provided by the Wellcome Trust (084743 to D.A.vH.); by grants from the Coeliac Disease Consortium, an Innovative Cluster approved by the Netherlands Genomics Initiative, and partially funded by the Dutch Government (BSIK03009 to C.W.) and the Netherlands Organisation for Scientific Research (NWO, VICI grant 918.66.620 to C.W.); by NIH grant 1R01CA141743 (to R.H.D); Fondo de Investigación Sanitaria FIS08/1676 and FIS07/0353 (to E.U.). This research utilizes resources provided by the Type 1 Diabetes Genetics Consortium, a collaborative clinical study sponsored by the National Institute of Diabetes and Digestive and Kidney Diseases, the National Institute of Allergy and Infectious Diseases, the National Human Genome Research Institute, the National Institute of Child Health and Human Development, and the Juvenile Diabetes Research Foundation International and is supported by NIH grant U01-DK062418. We acknowledge use of DNA from The UK Blood Services collection of Common Controls (UKBS-CC collection), funded by the Wellcome Trust grant 076113/C/04/Z and by NIHR programme grant to NHSBT (RP-PG-0310-1002). The collection was established as part of the Wellcome Trust Case Control Consortium (WTCCC)[33]. We acknowledge use of DNA from the British 1958 Birth Cohort collection, funded by the UK Medical Research Council grant G0000934 and the Wellcome Trust grant 068545/Z/02.

REFERENCES

1. Bingley PJ, et al. Undiagnosed coeliac disease at age seven: population based prospective birth cohort study. BMJ. 2004; 328:322–3. [PubMed: 14764493]
2. West J, et al. Seroprevalence, correlates, and characteristics of undetected coeliac disease in England. Gut. 2003; 52:960–5. [PubMed: 12801951]
3. van Heel DA, et al. A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. Nat Genet. 2007; 39:827–9. [PubMed: 17558408]

4. Hunt KA, et al. Newly identified genetic risk variants for celiac disease related to the immune response. Nat Genet. 2008; 40:395–402. [PubMed: 18311140]
5. Dubois PC, et al. Multiple common variants for celiac disease influencing immune gene expression. Nat Genet. 2010; 42:295–302. [PubMed: 20190752]
6. Trynka G, et al. Coeliac disease-associated risk variants in TNFAIP3 and REL implicate altered NF-kappaB signalling. Gut. 2009; 58:1078–83. [PubMed: 19240061]
7. Zhernakova A, van Diemen CC, Wijmenga C. Detecting shared pathogenesis from the shared genetics of immune-related diseases. Nature reviews. Genetics. 2009; 10:43–55. [PubMed: 19092835]
8. Smyth DJ, et al. Shared and distinct genetic variants in type 1 diabetes and celiac disease. N Engl J Med. 2008; 359:2767–77. [PubMed: 19073967]
9. Zhernakova A, et al. Meta-Analysis of Genome-Wide Association Studies in Celiac Disease and Rheumatoid Arthritis Identifies Fourteen Non-HLA Shared Loci. PLoS genetics. 2011; 7:e1002004. [PubMed: 21383967]
10. Cortes A, Brown MA. Promise and pitfalls of the Immunochip. Arthritis research & therapy. 2011; 13:101. [PubMed: 21345260]
11. Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–73. [PubMed: 20981092]
12. Clayton DG, et al. Population structure, differential bias and genomic control in a large-scale, case-control association study. Nature genetics. 2005; 37:1243–6. [PubMed: 16228001]
13. Galarneau G, et al. Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. Nature genetics. 2010; 42:1049–51. [PubMed: 21057501]
14. Spencer C, Hechter E, Vukcevic D, Donnelly P. Quantifying the underestimation of relative risks from genome-wide association studies. PLoS genetics. 2011; 7:e1001337. [PubMed: 21437273]
15. Lango Allen H, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010; 467:832–8. [PubMed: 20881960]
16. van Heel DA, Hunt K, Greco L, Wijmenga C. Genetics in coeliac disease. Best Pract Res Clin Gastroenterol. 2005; 19:323–39. [PubMed: 15925839]
17. Zhernakova A, et al. Evolutionary and functional analysis of celiac risk loci reveals SH2B3 as a protective factor against bacterial infection. Am J Hum Genet. 2010; 86:970–7. [PubMed: 20560212]
18. Holm H, et al. A rare variant in MYH6 is associated with high risk of sick sinus syndrome. Nature genetics. 2011
19. Lesage S, et al. CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. American journal of human genetics. 2002; 70:845–57. [PubMed: 11875755]
20. Johansen CT, et al. Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. Nature genetics. 2010; 42:684–7. [PubMed: 20657596]
21. Asimit J, Zeggini E. Rare variant association analysis methods for complex traits. Annual review of genetics. 2010; 44:293–308.
22. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare variants create synthetic genome-wide associations. PLoS biology. 2010; 8:e1000294. [PubMed: 20126254]
23. Zheng Q, Wang XJ. GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. Nucleic acids research. 2008; 36:W358–63. [PubMed: 18487275]
24. Lanktree MB, et al. Meta-analysis of Dense Genecentric Association Studies Reveals Common and Uncommon Variants Associated with Height. American journal of human genetics. 2011; 88:6–18. [PubMed: 21194676]
25. Donnelly P. Progress and challenges in genome-wide association studies in humans. Nature. 2008; 456:728–31. [PubMed: 19079049]
26. Lowe CE, et al. Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the IL2RA region in type 1 diabetes. Nature genetics. 2007; 39:1074–82. [PubMed: 17676041]

27. Genovese G, et al. Association of trypanolytic ApoL1 variants with kidney disease in African Americans. Science. 2010; 329:841–5. [PubMed: 20647424]
28. Shea J, et al. Comparing strategies to fine-map the association of common SNPs at chromosome 9p21 with type 2 diabetes and myocardial infarction. Nature genetics. 2011; 43:801–5. [PubMed: 21775993]
29. Jostins L, Morley KI, Barrett JC. Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets. European journal of human genetics : EJHG. 2011; 19:662–6. [PubMed: 21364697]
30. Asano A, Tsubomatsu K, Jung CG, Sasaki N, Agui T. A deletion mutation of the protein tyrosine phosphatase kappa (Ptprk) gene is responsible for T-helper immunodeficiency (thid) in the LEC rat. Mammalian genome : official journal of the International Mammalian Genome Society. 2007; 18:779–86. [PubMed: 17909891]
31. Adrianto I, et al. Association of a functional variant downstream of TNFAIP3 with systemic lupus erythematosus. Nature genetics. 2011; 43:253–8. [PubMed: 21336280]
32. Musunuru K, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature. 2010; 466:714–9. [PubMed: 20686566]
33. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447:661–78. [PubMed: 17554300]
34. Revised criteria for diagnosis of coeliac disease. Report of Working Group of European Society of Paediatric Gastroenterology and Nutrition. Archives of disease in childhood. 1990; 65:909–11. [PubMed: 2205160]
35. Romanos J, et al. Six new coeliac disease loci replicated in an Italian population confirm association with coeliac disease. Journal of medical genetics. 2009; 46:60–3. [PubMed: 18805825]
36. Plaza-Izurieta L, et al. Revisiting genome wide association studies (GWAS) in coeliac disease: replication study in Spanish population and expression analysis of candidate genes. Journal of medical genetics. 2011; 48:493–6. [PubMed: 21490378]
37. Megiorni F, et al. HLA-DQ and risk gradient for celiac disease. Human immunology. 2009; 70:55–9. [PubMed: 19027045]
38. Purcell S, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics. 2007; 81:559–75. [PubMed: 17701901]
39. Pruim RJ, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. 2010; 26:2336–7. [PubMed: 20634204]
40. Risch NJ. Searching for genetic determinants in the new millennium. Nature. 2000; 405:847–56. [PubMed: 10866211]

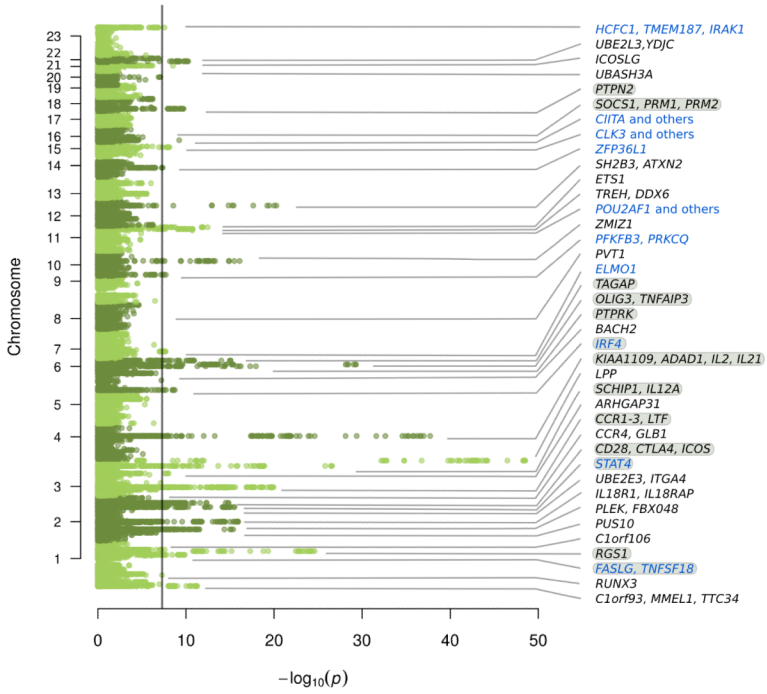


Figure 1. Manhattan plot of association statistics for known and novel celiac disease risk loci
Novel loci indicated in blue, loci with multiple signals indicated with grey highlight. Significance threshold drawn at $P=5\times10^{-8}$.

Europe PMC Funders Author Manuscripts

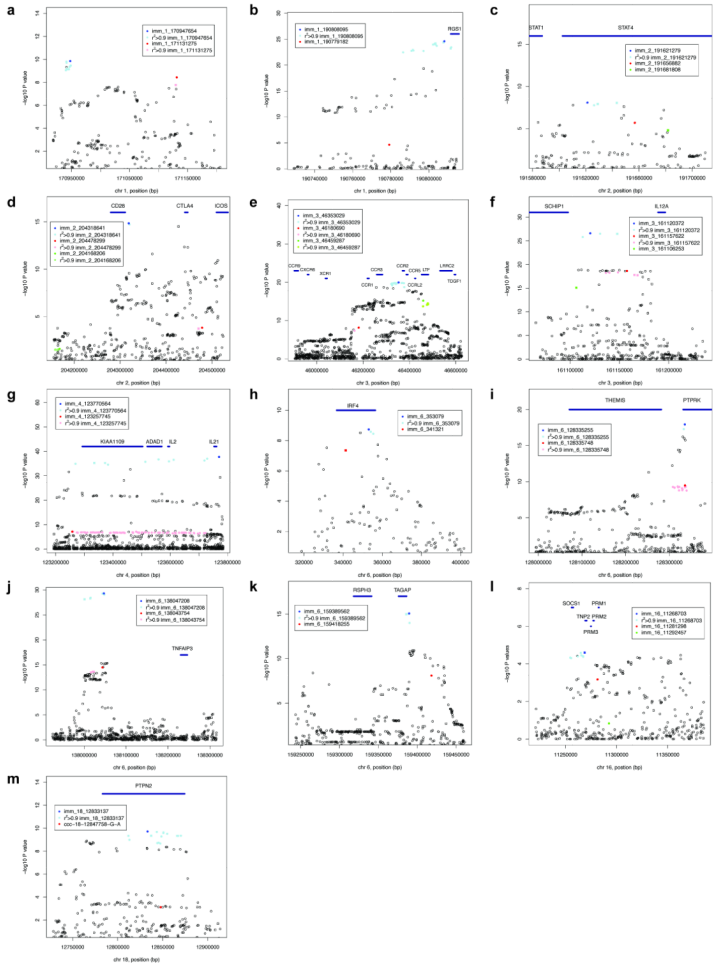


Figure 2. Loci with multiple independent signals

Non-conditioned P values shown for loci with multiple independent signals (from Table 2). The most associated variant for a signal shown in bold colour, further variants in r^2>0.90 (calculated from the 24,249 sample Immunochip dataset) shown in normal colour. First signal coloured blue, second coloured red, third coloured green. Squares indicate markers present in our previous celiac disease GWAS post quality control dataset (Illumina Hap550)[5].

Europe PMC Funders Author Manuscripts

Table 1

Sample Collections

Population sample	Celiac cases	Controls
UK	7728	8274[b]
The Netherlands	1123	1147
Poland	505	533
Spain - CEGEC[a]	545	308
Spain - Madrid[a]	537	320
Italy - Rome, Milan, Naples	1374	1255
India - Punjab	229	391
Total	12041	12228

[a]The two Spanish population samples were considered separately due to genotyping in different laboratories.

[b]5430 UK 1958 Birth Cohort participants, and 2844 UK Blood Services-Common Controls.

Each of the collections from the UK, Netherlands, Poland, Spain (Madrid) and Italy contained essentially the same sample set as our 2010 celiac disease GWAS[5], with now substantial additional samples from the UK and Netherlands and exclusion of amplified DNA samples from the Spanish collections. The Indian collection has not previously been studied. Our 2010 GWAS contained several collections not studied here.

Europe PMC Funders Author Manuscripts

Table 2

Risk variant signals at genome-wide significant celiac disease loci.

Non-HLA loci meeting genome-wide significance ($P<5\times10^{-8}$) in the current Immunochip dataset, or previous GWAS/replication dataset[5], are shown. Loci reported for the first time for celiac disease at genome wide significance are shown in bold in the Top variant column.

Top variant (dbSNP130 id)	Chr	HapMap3 CEU LD block[b] positions (hg18) (n markers, size)	MAF[c]	P[d]	OR	Highly correlated ($r^2>0.9$) variants positions (hg18) (n markers, size)	Localization: protein coding genes (RefSeq track UCSC/hg18)
rs4445406	1	2396747 - 2775531 (358, 379kb)	0.344	5.4×10^{-12}	0.87	2510162 - 2710035 (27, 200kb)	*C1orf93, MMEL1, TTC34*
rs72657048	1	25111876 - 25180863 (125, 69kb)	0.498	3.8×10^{-6}	0.92	25162321 - 25177139 (18, 15kb)	0 - 10kb 5′ & 1st exon *RUNX3*
rs12068671	1	170917308 - 171207073 (355, 290kb)	0.185	1.4×10^{-10}	0.86	170940206 - 170948695 (11, 8kb)	35 - 43kb 5′ FASLG
signal 2 rs12142280	1	”	0.180	8.3×10^{-9}*d*	0.87	171129607 – 171131275 (2, 2kb)	intergenic between *FASLG* and *TNFSF18*
rs1359062	1	190728935 - 190814664 (181, 86kb)	0.180	2.5×10^{-25}	0.77	190786488 - 190811722 (17, 25kb)	0 - 24kb 5′ & 1st exon *RGS1*
signal 2 rs72734930	1	”	**0.022**	3.7×10^{-4}*d*	1.23	190779182 (1)	32kb 5′ *RGS1*
rs10800746	1	199119734 - 199308949 (331, 189kb)	0.305	2.6×10^{-8}	0.89	199148015 (1)	9th intron *C1orf106*
rs13003464	2	60768233 - 61745913 (1047, 978kb)	0.388	4.3×10^{-16}	1.17	61040333 - 61058360 (3, 18kb)	exons 5-11 *PUS10*
rs10167650	2	68389757 - 68535760 (357, 146kb)	0.266	1.3×10^{-4}	0.92	68493221 - 68499064 (4, 6kb)	intergenic between *PLEK* and *FBX048*
rs990171	2	102221730 - 102573468 (894, 352kb)	0.225	1.2×10^{-16}	1.20	102338297 - 102459513 (45, 121kb)	*IL18R1, IL18RAP*
rs1018326	2	181502502 - 181972196 (898, 470kb)	0.418	3.1×10^{-16}	1.16	181708291 - 181803246 (24, 95kb)	intergenic between *UBE2E3* and *ITGA4*
rs6715106	2	191581798 - 191715979 (203, 134kb)	0.058	8.4×10^{-9}	0.79	191621279 - 191643278 (4, 22kb)	exons 6-14 *STAT4*
signal 2 rs6752770	2	”	0.296	1.3×10^{-6}*d*	1.10	191681808 (1)	intron 3 *STAT4*
signal 3 rs12998748	2	”	0.119	2.6×10^{-4}*d*	0.90	191656882 (1)	intron 3 *STAT4*
rs1980422	2	204154625 - 204524627 (642, 370kb)	0.233	1.4×10^{-15}	1.19	204318641 - 204320303 (2, 2kb)	intergenic between *CD28* and *CTLA4*
signal 2 rs34037980	2	”	0.217	1.6×10^{-5} *d*	0.91	204470572 – 204478299 (2, 8kb)	intergenic between *CTLA4* and *ICOS*

Europe PMC Funders Author Manuscripts

Top variant (dbSNP130 id)	Chr	HapMap3 CEU LD block[b] positions (hg18) (n markers, size)	MAF[c]	p[d]	OR	Highly correlated (r^2>0.9) variants positions (hg18) (n markers, size)	Localization: protein coding genes (RefSeq track UCSC/hg18)
signal 3 rs10207814	2	”	**0.039**	1.3×10^{-4} *d*	1.20	204158521 - 204168206 (5, 10kb)	111 – 121 kb 5′ *CD28*
rs4678523	3	32895606 - 33063377 (260, 168 kb)	0.313	2.4×10^{-7}	1.11	33012725 - 33012756 (2, 31bp)	intergenic between *CCR4* and *GLB1*
rs2097282	3	45904804 - 46625997 (1343, 721kb)	0.314	1.1×10^{-20}	1.20	46321275 - 46377631 (27, 56kb)	intergenic between *CCR3* and *CCR2*
signal 2 rs7616215	3	”	0.361	8.6×10^{-9} *d*	1.12	46162711 – 46180690 (2, 18kb)	38 – 55 kb 3′ CCR1
signal 3 rs60215663	3	”	0.070	4.8×10^{-5} *d*	1.16	46458634 – 46480319 (7, 22kb)	exons 2-13 *LTF* (NM_002343.3)
rs61579022	3	120587671 - 120783345 (372, 196kb)	0.390	9.9×10^{-9}	1.11	120601187 - 120605968 (4, 5kb)	intron 10 ARHGAP31
[imm_3_161120372]	3	161065075 - 161237201 (423, 168kb)	0.111	2.6×10^{-27}	1.36	161112778 - 161147744 (4, 35kb)	intergenic between *SCHIP1* and *IL12A*
signal 2 rs1353248	3	”	0.288	9.8×10^{-9} *d*	0.88	161106253 (1)	intergenic between *SCHIP1* and *IL12A*
signal 3 rs2561288	3	”	0.455	8.1×10^{-8} *d*	1.12	161136316 – 161168494 (6, 32kb)	intergenic between *SCHIP1* and *IL12A*
rs2030519	3	189552054 - 189622323 (142, 70kb)	0.486	3.0×10^{-49}	0.76	189587750 - 189602595 (8, 15kb)	intron 2 *LPP*
rs13132308	4	123192512 - 123784752 (1294, 592kb)	0.166	1.9×10^{-38}	0.71	123269042 - 123770564 (11, 502kb)	multiple genes *(KIAA1109, ADAD1, IL2, IL21)*
signal 2 rs62323881	4	”	0.073	8.6×10^{-5} *d*	1.15	123257527 – 123722990 (87, 465kb)	multiple genes *(KIAA1109, ADAD1, IL2, IL21)*
rs1050976	6	315547 - 402748 (199, 87kb)	0.488	1.8×10^{-9}	0.89	353079 - 355417 (3, 2kb)	3′ UTR *IRF4* (NM_002460.3)
signal 2 rs12203592	6	”	0.183	2.6×10^{-4} *d*	0.91	341321 (1)	intron 4 *IRF4* (NM_002460.3)
rs7753008	6	90863556 - 91096529 (341, 233kb)	0.380	2.7×10^{-7}	1.10	90866360 - 90875874 (5, 10kb)	intron 2 *BACH2* (NM_001170794.1)
rs55743914	6	127993875 - 128382483 (572, 389kb)	0.239	1.1×10^{-18}	1.21	128332892 - 128335255 (2, 2kb)	*PTPRK* last exon, 3′UTR (NM_002844.3)
signal 2 rs72975916	6	”	0.150	1.2×10^{-5} *d*	0.89	128307943 - 128339304 (15, 31kb)	*PTPRK* exons 28-30, 3′UTR, to 24kb 3′
rs17264332	6	137924568 - 138316778 (864, 392kb)	0.211	5.0×10^{-30}	1.29	138000928 - 138048197 (6, 47kb)	intergenic between *OLIG3* and *TNFAIP3*
signal 2 [imm_6_138043754]	6	”	0.190	2.1×10^{-7} *d*	0.88	138015797 – 138043754 (4, 28kb)	intergenic between *OLIG3* and *TNFAIP3*

Europe PMC Funders Author Manuscripts

Top variant (dbSNP130 id)	Chr	HapMap3 CEU LD block[b] positions (hg18) (n markers, size)	MAF[c]	p[d]	OR	Highly correlated (r^2>0.9) variants positions (hg18) (n markers, size)	Localization: protein coding genes (RefSeq track UCSC/hg18)
rs182429	6	159242314 - 159461818 (514, 220kb)	0.427	8.5×10^{-16}	1.16	159385965 - 159390046 (4, 4kb)	4kb 5′ and 5′ UTR *TAGAP* (NM_152133.1)
signal 2 rs1107943	6	”	0.071	2.8×10^{-6} *d*	1.18	159418255 (1)	32kb 5′ *TAGAP* (NM_152133.1)
[1kg_7_37384979]	7	37330503 - 37406978 (213, 76kb)	0.101	2.1×10^{-8}	1.18	37366994 - 37404402 (31, 37kb)	intron 1 *ELMO1*
rs10808568	8	129211716 - 129368419 (400,157kb)	0.256	2.2×10^{-5}	0.91	129333242 - 129345888 (4, 13kb)	151 - 163kb 3′ of *PVT1*
rs2387397	10	6428077 - 6585110 (411, 157kb)	0.229	1.9×10^{-8}	0.88	6430198 (1)	intergenic between *PFKFB3* and *PRKCQ*
rs1250552	10	80690408 - 80774414 (223, 84kb)	0.470	8.0×10^{-17}	0.86	80728033 (1)	intron 14 *ZMIZ1*
rs7104791	11	110682429 - 110815769 (3, 133kb)	0.209	1.9×10^{-11}	1.16	not high-density genotyped	[region: *POU2AF1*, *C11orf93*]
rs10892258	11	117847131 - 118270810 (466, 424kb)	0.237	1.7×10^{-11}	0.86	118080536 - 118085075 (5, 5kb)	intergenic between *TREH* and *DDX6*
rs61907765	11	127754640 - 127985723 (480, 231kb)	0.213	3.4×10^{-13}	1.18	127886184 - 127901948 (6, 16kb)	5kb 5′ & 1st exon *ETS1* (NM_001162422.1)
rs3184504	12	110183529 - 111514870 (938, 1331kb)	0.488	5.4×10^{-21}	1.19	110368991 - 110492139 (4, 123kb)	5′ UTR & exons 1-3 *SH2B3*; exons 2-25 & 3′ UTR *ATXN2*
rs11851414	14	68238574 - 68387815 (338, 149kb)	0.221	4.7×10^{-8}	1.13	68329159 - 68341722 (3, 13kb)	1kb 5′ & 1st exon *ZFP36L1*
rs1378938	15	72397784 - 73270664 (23, 873kb)	0.278	7.8×10^{-9}	1.13	not high-density genotyped	[region inc. *CLK3, CSK* and multiple genes]
rs6498114	16	10834038 - 10903351 (8, 69kb)	0.246	5.8×10^{-10}	1.14	not high-density genotyped	[region: *CIITA*]
rs243323	16	11220552 - 11385420 (446, 165kb)	0.300	2.5×10^{-5}	0.92	11254549 - 11268703 (12, 14kb)	11kb 5′, all of SOCS1, 1kb 3′
signal 2 [imm_16_11281298]	16	”	**0.004**	1.3×10^{-4} *d*	1.70	11281298 (1)	intergenic between *PRM1* and *PRM2*
signal 3 rs9673543	16	”	0.169	2.0×10^{-4} *d*	1.10	11292457 (1)	10kb 5′ PRM1
rs11875687	18	12728413 - 12914117 (411, 186kb)	0.150	1.9×10^{-10}	1.17	12811903 - 12870206 (16, 58kb)	exons 2-5 *PTPN2* (NM_080422.1)
signal 2 rs62097857	18	”	**0.040**	5.2×10^{-5} *d*	1.20	12847758 (1)	intron 2 *PTPN2* (NM_080422.1)
rs1893592	21	42683153 - 42760214 (226, 77kb)	0.282	3.0×10^{-9}	0.88	42728136 (1)	intron 9 *UBASH3A* (NM_018961)

Europe PMC Funders Author Manuscripts

Top variant (dbSNP130 id)	Chr	HapMap3 CEU LD block[b] positions (hg18) (n markers, size)	MAF[c]	p[d]	OR	Highly correlated (r^2>0.9) variants positions (hg18) (n markers, size)	Localization: protein coding genes (RefSeq track UCSC/hg18)
rs58911644	21	44414408 - 44528088 (239, 114kb)	0.193	6.2×10^{-7}	0.89	44446245 - 44453549 (8, 7kb)	18 - 25kb 3′ *ICOSLG*
rs4821124	22	20042414 - 20352005 (131, 310kb)	0.186	5.7×10^{-11}	1.16	20250903 - 20313260 (36, 62kb)	*UBE2L3, YDJC*
rs13397	X	152825373 - 153043675 (88, 218kb)	0.133	2.7×10^{-8}	1.18	152872114 - 152937386 (4, 65kb)	*HCFC1, TMEM187, IRAK1*

[a] Only the most significantly associated risk variant from each region and independent signal is shown. Variant names shown are as in dbSNP130 where available. Otherwise, the Illumina Immunochip manifest name is shown in brackets (Supplementary Table 5 shows both names for variants).

[b] Regions were first defined by linkage disequilibrium blocks, extending 0.1 cM to the left and right of the risk SNP as defined by the HapMap3 CEU recombination map. For loci with multiple different previously reported risk SNPs for different diseases, and overlapping blocks, the extended region is shown. Regions where additional case resequencing (as well as 1000Genomes) has been performed are shown, with boundaries of the resequencing effort(s). All chromosomal positions are based on NCBI build-36 (hg18) coordinates.

[c] MAF shown for European controls. See Supplementary Table 4 for more detailed allele frequencies in cases and controls by collection. Low frequency and rare variants shown in bold.

[d] Logistic regression association test. Tests for second (and third) independent signals are conditioned on the first (and second) reported variant(s). Per locus significance thresholds for second (and third) independent signals are shown in Supplementary Table 3.

Molecular Genetics of Coeliac Disease

Vanisha Mistry, *Blizard Institute, Queen Mary University of London, London, UK*

David van Heel, *Blizard Institute, Queen Mary University of London, London, UK*

Advanced article

Article Contents

Online posting date: 15th December 2011

Coeliac disease (CD) is a common inflammatory disease of the small intestine. It has a prevalence of 1% in the population and is strongly heritable. Current germline disease risk variants explain ~50% of known heritability, the majority contributed by a strong human leukocyte antigen (HLA)-DQ association. The role of HLA-DQ in the immunology of CD is well understood, for example the role of tissue transglutaminase and HLA-DQ in modifying and binding immuno-dominant dietary cereal (gluten) peptides. Genome-wide association studies have found 39 loci with risk variants of more modest effect. The use of high-throughput sequencing technologies to locate rare variants of larger effect may aid in the complete resolution of this complex trait, as well as in other autoimmune diseases, which show considerable overlap in immunological pathways.

Introduction

Coeliac disease (CD) (or gluten sensitive enteropathy) is a common autoimmune disorder of the small intestine. An immune response is generated by the presence of gluten – storage proteins found in wheat, barley and rye – in genetically susceptible individuals, triggering an inflammatory response causing intestinal morphological changes. Although CD shares similarities in its symptoms with other inflammatory bowel diseases, it is the ingestion of gluten that elicits an abnormal immune response in susceptible individuals. Its molecular basis can be described as a quantitative polygenic trait as the outcome phenotype is consequential of combinations of genes on multiple loci having an effect on each other. The identification of the human leukocyte antigen (HLA) DQ gene variants and their role in CD has contributed to our understanding of disease. Recent studies have implicated other disease loci mostly located in immunological pathways.

eLS subject area: Genetics & Disease

How to cite:
Mistry, Vanisha; and van Heel, David (December 2011) Molecular Genetics of Coeliac Disease. In: eLS. John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0022476

Clinical Manifestations and Pathophysiology

CD affects the mucosa of the small intestine, leading to presentation of symptoms such as malabsorption, malnutrion, steatorrhea (diarrhoea caused by excess fat), weight loss, abdominal pain and anaemia (**Figure 1**). The mucosa of the small intestine is covered by villi: finger-like projections with a large surface area for absorption. Its core is an extension of the lamina propia and crypts (circular intestinal cells) lie at the base. The dietary prolamins – storage proteins in grain – trigger inflammation. One gluten protein is gliadin; these peptides pass through the epithelial barrier of the intestine into the lamina propria, and deamidation occurs due to generation of autoantibodies to the enzyme tissue transglutaminase (TG2) (Molberg *et al.*, 1998; **Figure 2**). This increases peptide affinity to HLA class II molecules (HLA-DQ2 or HLA-DQ8), generating $CD4^+$ T-helper 1 cell (Th1) activation. Intraepithelial lymphocytes (IELs) have the role of destroying epithelial cells in the mucosa via natural killer receptors (NKRs) expressed on their surface. It is through NKRs that IELs recognise MHC-1 molecules induced on the surface of enterocytes by stress and inflammation. Here, armed effector IELs are activated to lymphokine-activated killing cells, producing epithelial cell death in T cell receptor – independent manner. CD is described as a Th1-mediated disease, as it shows up-regulation of interferon (IFN)-γ production and T-bet levels in gut infiltrating cells (Holtmann and Neurath, 2004). An increased number of T cells are prevalent in individuals with CD; upon gluten ingestion there is infiltration of T cells in the lamina propria followed by crypt hyperplasia and villous atrophy.

Epidemiology

Although gluten is one environmental factor (other factors are likely but yet unidentified), there are multiple inherited genetic factors affecting disease susceptibility inferred from

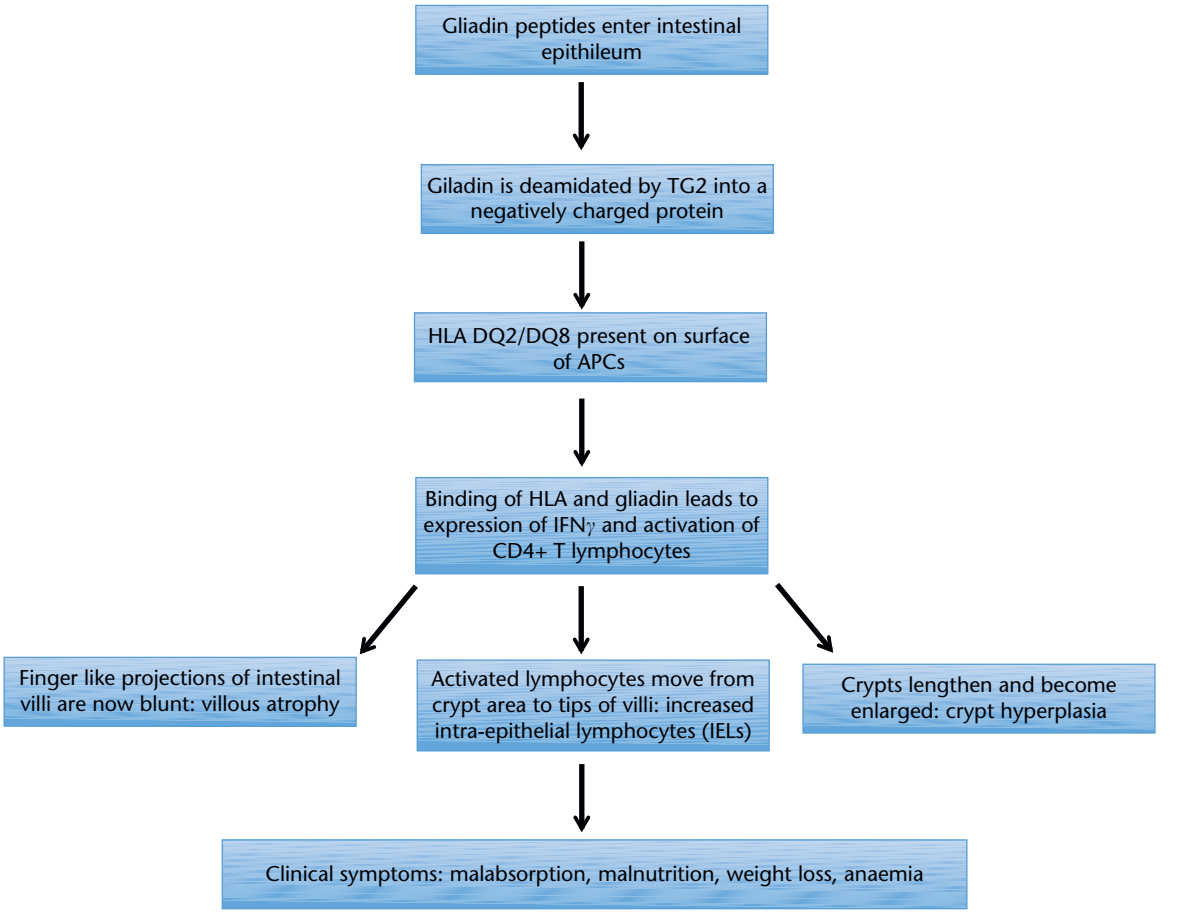


Figure 1 Flow diagram highlighting clinical manifestations of disease upon gluten ingestion.

family and twin studies. The sibling recurrence risk ratio is between 10 and 30 for disease development, and one of a dizygotic twin pair has a 70% chance of developing disease (Greco *et al*., 2002). Approximately 1 in 100 individuals of European descent and .4–.95% of individuals in USA (Dube *et al*., 2005) have CD, but it is less common in Asia and South America.

The dispersion of HLA-DQ variants coincides with disease occurrence; HLA-DQ8 is most common in Latin America and Northern Europe, whereas HLA-DQ2 is most common in Western Europe, North and West Africa, the Middle East and Central Asia (Cummins and Roberts-Thomson, 2009). High frequencies of HLA-DQ2 are present in the Saharawi population of Algeria, where the prevalence of disease is 5.6% (Catassi *et al*., 1999), contrasting to almost negligible prevalence in the Chinese/Japanese population.

The ratio of diagnosed to undiagnosed disease is 1:7 (Heap and van Heel, 2009); testing for disease presence has improved with more sensitive and specific serological screenings. Testing for the presence of immunoglobulin A (IgA) autoantibodies to endomysium is a specific marker for the presence of CD (Dieterich *et al*., 1998). If positive, confirmation is necessary by biopsy of the small intestine. The Marsh classification is used systematically for diagnosis according to small bowel pathology. HLA typing can be useful for exclusion in patients with equivocal histological finding but has low specificity (Kaukinen *et al*., 2002).

Currently adhering to a gluten-free diet (GFD) leads to disease remission and symptoms typically reduce within a few weeks. There are subsets of patients who do not respond to GFD (2–5%), mainly those diagnosed at age of 50 or above; this is known as refractory CD. The leading cause of death for these patients is enteropathy-associated T cell lymphoma (Al-toma *et al*., 2007).

Immunogenetics in CD

CD, similar to other autoimmune diseases, shows strong association with major histocompatibility complex (MHC)-associated genes, which play a role in the immune system. These molecules are encoded at chromosome 6p21; MHC class II molecules are encoded at gene loci HLA-DQ, HLA-DP and HLA-DR. **See also**: Major Histocompatibility Complex: Disease Associations

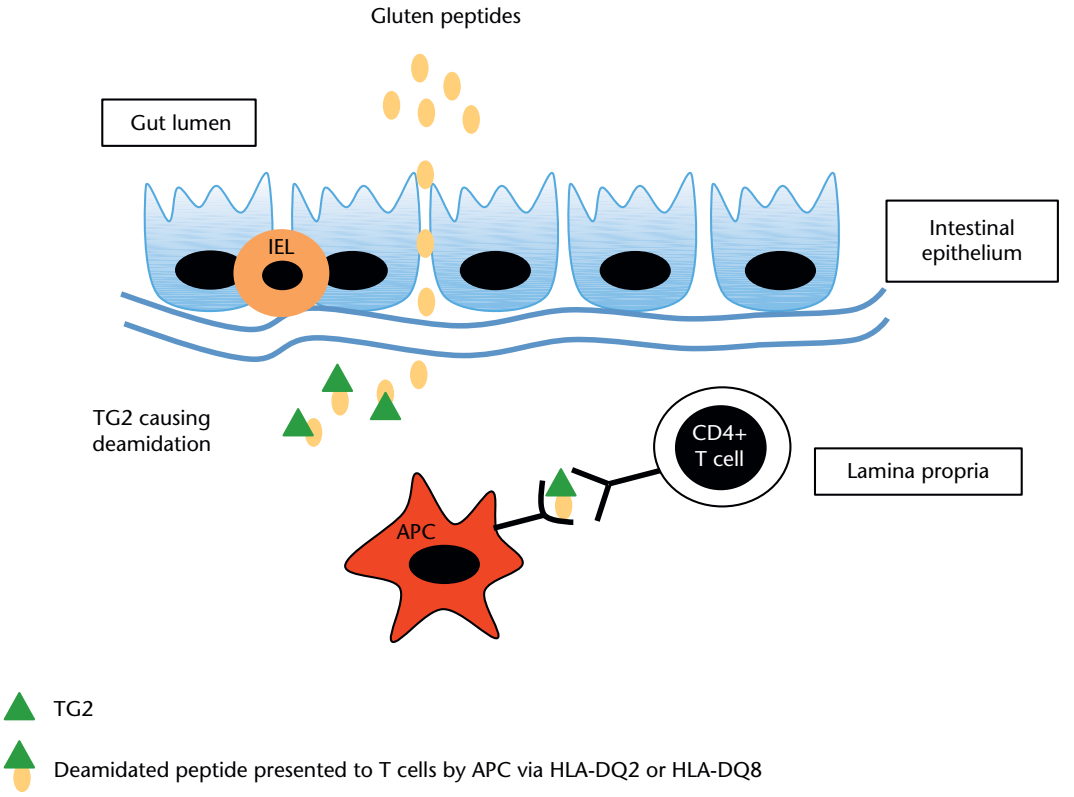


Figure 2 Model of deamidated gluten peptide presentation by APC to T cells for subsequent loading onto HLA-DQ2 or HLA-DQ8 heterodimers. Gluten peptides (or gliadin peptides, which is a gluten protein) pass through a fairly permeable epithelial layer of the small intestine in untreated coeliac disease. Intestinal permeability is compromised by IELs producing more interferon thereby intensifying the immune reaction. Gliadin peptides react with transglutaminase 2 (TG2) serum autoantobodies in the lamina propria. TG2 is the autoantigen of coeliac disease and plays a primary role of crosslinking and deamidation of gliadin. Ingested gliadin is crosslinked by TG2 causing specific deamidation of glutamine into glutamic acid. After deamidation the gliadin peptides can be presented more efficiently to gliadin reactive CD4 T cells by APCs via HLA-DQ2 or HLA-DQ8.

HLA association

The most common genetic background coeliac individuals share is the presence of either HLA-DQ2 or HLA-DQ8 serotype present on the HLA-DQ $\alpha\beta$ heterodimer (**Figure 3**). These molecules are expressed on antigen-presenting cells (B cells, dendritic cells and macrophages). The early identified HLA-DQ2 serotype was found to be primarily associated with disease (Tosi *et al.*, 1983), which is mediated through the DQ2.5 haplotype; the DQ2.2 haplotype does not appear to predispose to disease (Sollid *et al.*, 1989). HLA-DQ2 is encoded by HLA-DQA1∗05 allele (alpha chain) and HLA-DQB1∗02 allele (beta chain). The two alleles are present in *cis* conformation on the DR3 haplotype. 90% of European patients carry the HLA-DQ2 heterodimer and the remaining carries either one DQ2 allele or HLA-DQ8 (Karell *et al.*, 2003). HLA-DQ8 is encoded by HLA-DQA1∗03 (alpha chain) and HLA-DQB1∗0302 alleles (beta chain).

It is possible to generate a combination of DQ2.2 and DQ2.5 haplotypes depending on parental genotype since each haplotype is only present on one chromosome. If in *cis* conformation, both alpha and beta chain are encoded on the same chromosome rather than each parent supplying one chain (*trans* conformation). One or two copies of HLA-DQ2 give an intermediate or high risk for disease. Monsuur *et al.* (2008) predicted HLA-associated risk factors using tagging SNPs. They showed increased risk individuals were homozygous for the DQ2.5 haplotype or possessed a single copy of DQ2.5 and one copy of DQ2.2, DQ2.7 or DQ2.8 (Monsuur *et al.*, 2008). This coincides with previous findings by van Belzen *et al.* (2004) who reported that being homozygote for DQ2.5 gives a 4–6 times increased risk of disease (van Belzen *et al.*, 2004).

The presence of HLA class II genes is not the sole genetic component for disease; HLA-DQ2 is expressed in 30% of the European population (Sollid *et al.*, 1989), with 2–5% of gene carriers developing disease. These early findings suggest other genetic factors contribute to the manifestation of CD.

Finding Disease Genes

To determine absolute risk of disease, non-HLA risk alleles in CD must be taken into account. The past decade has been subject to a series of successful genetic studies all with the aim of finding risk variants susceptible to common

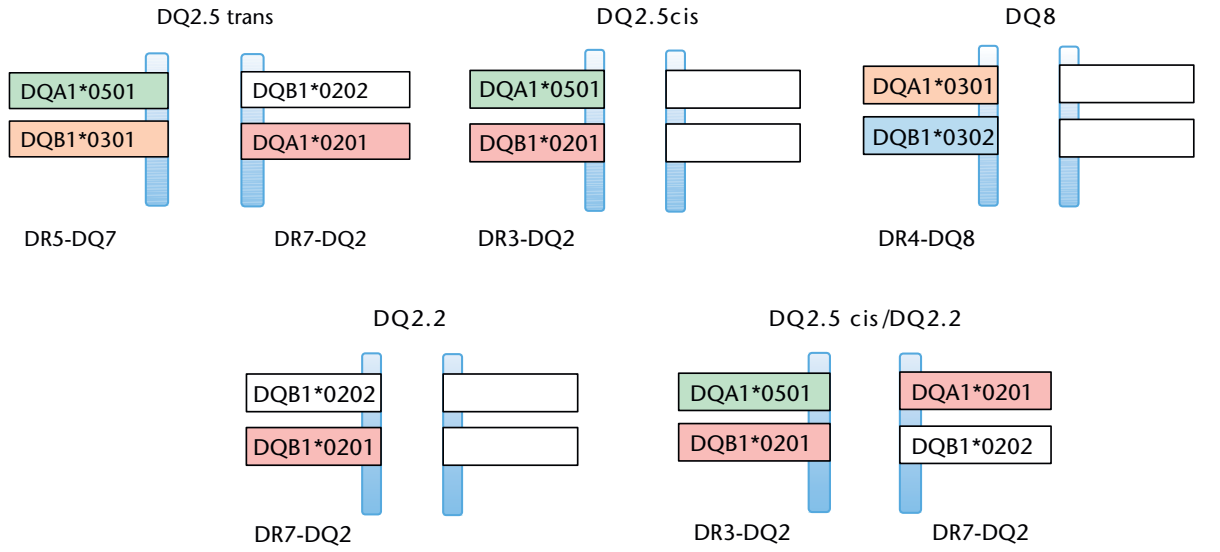


Figure 3 HLA haplotype combinations in coeliac disease. White boxes denote 'other' haplotype. DQ2.5 *cis* is shown as a heterozygote; a DQ2.5 *cis* homozygote will carry same alleles on both chromosomes. Majority of CD patients express HLA-DQ2.5 encoded either in *cis* on the DR3-DQ2 haplotype, or in *trans* on the DR5-DQ7/DR7-DQ2 haplotype for heterozygous individuals. HLA-DQ2.2 confers low risk for CD if expressed solely. HLA-DQ8 is expressed in DQ2-negative patients (Abadie *et al.*, 2011). Adapted from Dubois and van Heel (2008).

complex disease. The completion of the Human Genome Project in 2003 allowed significant progress in family linkage studies; genetic markers spanning the entire human genome enabled extensive mapping efforts resulting in the discovery of many genes for Mendelian diseases and traits. Recently, the mapping of common and low frequency variation (International HapMap Consortium, 1000 Genomes Project Consortium) has made it possible to locate and map genes surrounding disease risk loci by genome-wide association studies (GWAS).

Linkage disequilibrium

The concept of linkage disequilibrium (LD) is centralised on the nonrandom association of alleles at different loci. A SNP occurs once every 100–300 bases in the genome. Natural selection, or chance, caused the spread of common SNP mutations that arose thousands of generations ago; a second mutation occurring later but close to an earlier one results in both alleles being transmitted to the same offspring in subsequent generations. It is this model that is exploited in GWA studies. An increased risk of disease caused by one SNP denotes direct association between that SNP and disease in the population and indirect association between several nearby SNPs, due to LD. Therefore it is possible to identify association in the chromosomal region without genotyping every SNP in the genome.

International HapMap Project and 1000 Genomes Project

The International HapMap Project commenced in 2003 with a focus to map all common genetic variation (greater than 5% minor allele frequency) across several populations, equating to 3.5 million SNPs. Approximately 90% of the genetic variation in the Caucasian population has been captured. A valuable outcome is the determination of LD in four major populations, confirming its usefulness as a reference sample in GWA investigations.

The 1000 Genomes Project (1000G) was set up in 2007 to capture low frequency variants (1%) in the human genome (http://www.1000genomes.org). The pilot phase included low coverage sequencing of 179 individuals from four populations, high coverage sequencing of two trios and exon sequencing of nearly 700 individuals from seven populations (Durbin *et al.*, 2010). This public reference catalogue of human genetic variation will aid in more GWAS identifying previously missed associations and provide a filter in Mendelian disease for exclusionary purposes.

Genetic linkage studies

Family-based designs have been used in genetic studies since Mendel's laws of inheritance dominated the fundamental concepts of genetics. The independence of segregation, as inferred by Mendel's law of segregation, is not always true: there are group of traits that are linked and the genes controlling them tend to be inherited together by the offspring as a group, not independently. Heritable variation is reliant on the outcome of meiotic transmission and linkage analysis is the inference of the outcome of meiosis. The underlying principle is that if two individuals are phenotypically similar that is carry disease, then a genetic marker located near a disease susceptibility gene must also be similar that is shared by both carriers. The factors contributing to the linkage model, which will prove or disprove the null of no linkage, are the overall contribution

of the trait loci and the genetic distance between the disease gene and marker being tested. **See also**: Genetic Linkage Mapping

Genome-wide association studies

GWAS using SNP markers is a more powerful approach for elucidating genetic determinants than family-based linkage studies. It was developed in tandem with the 'common disease common variant' hypothesis, recognising that multiple genomic loci were likely to be involved in susceptibility to common multifactorial traits due to variants being present at relatively high frequency with an individually small magnitude of effect. **See also**: Genome-wide Association Studies: The Success, Failure and Future

Some studies have been follow-ups from linkage signals or candidate gene studies to narrow down association to a single haplotype, such as the region near the *CTLA4* gene in CD (Hunt *et al*., 2005). SNP markers across the whole genome are tested for association with a disease in a large cohort of disease cases compared with a similar or higher number of controls. After performing correctional tests common variants in correlation with disease are identified depending on the risk allele frequency, its association between marker genotyped and relative risk conferred by genotype. Up to now associated variants have been found mostly in noncoding regions by GWAS, so it is accepted that common variant contribution to disease is more likely to be of regulatory function rather than protein coding.

Known Genetic Structure of CD

As noted, the most significant association to CD is with HLA-DQ2. Possession of HLA-DQ2 serotypes is necessary for affinity to deamidated gliadin, yet 30% of the Caucasian population also carry HLA-DQ2 without developing disease. Despite its importance in disease pathophysiology HLA contribution to disease is currently estimated at approximately 40% (Dubois *et al*., 2010).

Regions identified through linkage analysis

Unlike Mendelian disease, complex disease has had less success in finding causal variants through linkage. A large number of non-HLA variants have been implicated in CD yet show lack of significant association in multiple populations. Linkage was found to various regions including 5q (Percopo *et al*., 2003) and 19p (van Belzen *et al*., 2003); the former was replicated in a meta analysis of multiple populations (Babron *et al*., 2003). One region containing *CTLA4*, *ICOS* and *CD28* is involved in immune suppression hence was highlighted as a valid candidate. Haplotype analysis showed strong association in the Irish population (Brophy *et al*., 2006) and variants in the 3′ region of *CTLA4* were thought to influence responses in type 1 diabetes (T1D) (King *et al*., 2000). In spite of promising initial analysis, replication in genome-wide scans was inconsistent for this region (King *et al*., 2003).

The power of family studies is decreased due to small effect sizes attributable to genetic variants present at high frequency. An exception is *NOD2* in Crohn's disease (Hugot *et al*., 2001) and HLA replication in CD due to common encoding variants being of large effect size, hence having sufficient power. This explains that predisposition to complex disease is not caused by just a handful of highly penetrant mutations but a mixture of multiple risk variants with varying effect size.

Susceptibility gene loci identified by GWAS

To date two GWA studies in the UK have been carried out in CD identifying non-HLA variants. In the first study 778 coeliac cases and 1422 matched population controls using 310605 SNPs showed highest association on chromosome 4q27 harbouring the *KIAA1109-TENR-IL2-IL21* LD block (van Heel *et al*., 2007). Follow up studies found associations in *REL*, *TNFAIP3* (Trynka *et al*., 2009) and a region encompassing *CTLA4*, *ICOS* and *CD28* (Smyth *et al*., 2008). The UK follow up replication study by Hunt *et al*. (2008) identified a further seven regions meeting genome-wide combined significance ($p = < 5 \times 10^{-7}$) (Hunt *et al*., 2008). An additional association was found in *ITGA4* in a US case control collection (Garner *et al*., 2009). The second generation GWAS by Dubois *et al*. (2010) recognised a further 13 genome-wide significant regions, most containing genes controlling immune responses. This study estimated current non-HLA loci to account for 6% of the total genetic variance of CD, whereas the rest is dominated by HLA contribution increasing the heritability estimate to just under 50% (Dubois *et al*., 2010).

Associated CD loci

Current associated loci point to primary altering of the immune system response in several immune-related pathways. Three main associated loci reported before and replicated in the second GWAS (Dubois *et al*., 2010), are briefly discussed.

IL2/*IL21* region

This region is the strongest non-HLA marker. It is contained in the ~700 kb LD block of the 4q27 chromosomal region. Recently significant associations between the SNP rs6822844, located in the intergenic *IL2*/*IL21* region, and six autoimmune diseases have been reported (Maiti *et al*., 2010) implicating it as a general autoimmune disease susceptibility locus.

IL2 is involved in stimulating T cell activation and proliferation, but can also stimulate proliferation of natural killer (NK) cells and immunoglobin production from B cells. It maintains CD4 and CD25 T_{reg} cells and destroys self-reactive T cells via activation-induced cell death (Fontenot *et al*., 2005).

IL21 acts on NK cells, $CD4^+$ T cells and B lymphocytes to induce and sustain antibody production after tissue

damage (Sarra *et al*., 2011). This region has been shown to induce messenger ribonucleic acid synthesis for genes involved in activating innate immunity and Th1 response, such as T-bet and IFN-gamma (Strengell *et al*., 2002). There is evidence that *IL21* is increased in untreated coeliac mucosa (Caruso *et al*., 2007).

RSG1

RGS1 regulates activity in G-protein signalling and is involved in B cell proliferation and activation. Interestingly, it shows expression in IELs (Hunt *et al*., 2008), which recognise MHC-1 molecules through NK receptors in CD. *RGS1* also regulates chemokine receptor signalling and B cell trafficking to lymph nodes in mice (Han *et al*., 2005). Recently, Tran *et al*. (2010) illustrated how IFN-β can induce the expression of *RGS1* in peripheral blood cells of MS patients, suggesting involvement in disease treatment (Tran *et al*., 2010).

SH2B3

SH2B3 is an adapter protein with pleckstrin homology and is involved in inhibiting T cell inactivation via Src homology 2 domains (Li *et al*., 2000). It mediates interaction between T cell receptors and intracellular signalling pathways. It also inhibits the activation of nuclear factor on activated T cells, which binds to deoxyribonucleic acid regulating the expression of cytokines, including *IL2*. *SH2B3* has been implicated in having a protective role against bacteria infection in CD as carriers of the risk allele, rs3184504∗A, have strong activation of the *NOD2* recognition pathway (Zhernakova *et al*., 2010).

Overlap with other autoimmune disease

The numbers of autoimmune disease loci that overlap with CD highlight shared immunological pathways, mainly resulting in inactivation of T cells. Approximately 64% of 39 known CD loci are shared with at least one other autoimmune disease (Gutierrez-Achury *et al*., 2011). The main shared loci are with T1D, rheumatoid arthritis, Crohn's disease and ulcerative colitis (**Table 1**).

Finding causal variants

GWAS findings reveal associated variants that are mostly common, have modest to weak effect sizes, and are credible disease markers, however only explain a small percentage of heritability. Subsequent to discovering a region harbouring a risk locus, all genetic variants in the locus require detailed analysis to find the causative variant and to establish its quantitative contribution to disease. Follow up *in vivo* studies are necessary to investigate functional pathways and deduce biological mechanisms of disease susceptibility or resistance.

Pinpointing causal variants requires fine mapping of the associated regions(s), followed by targeted resequencing, exome (exons of genes) or whole genome sequencing. These techniques are becoming common practice due to ongoing advances in high-throughput sequencing technologies. Attention is now being focused on low frequency (.5–5% minor allele frequency) and rare variant (below .5%) analysis, shown to contribute significantly to genetic architecture of disease (Durbin *et al*., 2010) coinciding with the 'rare variant common disease' hypothesis.

Fine mapping

Fine mapping is a necessary step after genotyping to refine associated region(s) to a causal variant(s) by analysing a high density of genetic markers across the LD region. Fine mapping in coeliac-associated regions are yet to be published, however extensive mapping across the MHC region for discovery of HLA-linked loci to T1D has been carried out (Brown *et al*., 2009). Mapping around the 4 Mb region established HLA-B and HLA-A to be associated independently of HLA class II genes in TID, a contrast to CD where no other independent associations in the HLA region have been found. Results from this study iterated the multilocus effects due to classical HLA genes and extensive LD spanning the entire region. The major susceptibility gene to T1D was refined to two independent groups of SNPs encompassing *IL2RA* intron 1 and '5' regions of *IL2RA* and *RBM17* after large-scale fine mapping (Lowe *et al*., 2007). Present research in SLE has localised the effect of *IL2*/*IL21* locus to two SNPs in high LD through fine mapping of 45 tag SNPs in the region (Hughes *et al*., 2011). Furthermore, genome-wide imputation from a reference panel (i.e. CEU HapMap) can show much stronger associations; this has been highlighted for the *TAGAP* risk locus in rheumatoid arthritis (Plenge *et al*., 2010; Chen *et al*., 2011).

At present, custom-designed platforms are being used in association studies of several chronic and autoimmune disease loci (Immunochip) and cardiovascular disease, type 2 diabetes and obesity loci (Metabo-Chip) to fine map-associated variants to smaller regions and search for rare variation to assist in closing the heritability gap.

Targeted resequencing, exome and whole genome sequencing

It has been frequently proposed that rare mutations of larger effect size account for a substantial proportion of the missing heritability in disease (Pritchard, 2001). Identifying rare mutations altering protein function independent of common SNPs is the hopeful outcome of targeted resequencing in an associated candidate region. Momozawa *et al*. (2011) identified low frequency coding variants from 63 GWAS-identified positional candidate genes by showing protection against inflammatory bowel disease in *IL32R* but found no rare variants predisposed to Crohn's disease (Momozawa *et al*., 2011). In contrast, four rare variants in the *IFIH1* gene were found to confer protection against T1D by altering protein expression and structure (Nejentsev *et al*., 2009).

Another approach is to sequence the exomes of disease individuals to capture functional variation ($\sim$30 Mb) predisposing to disease (Ng *et al*., 2008). Protein coding

Table 1 39 non-HLA coeliac loci showing association with other autoimmune diseases

Associated coeliac loci	Reported genes	Overlapping autoimmune diseases[d]	References
1q24.2[c]	*CD247*	Rheumatoid arthritis	Plenge *et al*. (2010)
1q24.3[c]	*FASLG, TNFSF18, TNFSF4*	Crohn's disease	Barrett *et al*. (2008)
1q32.1[b]	Intergenic	Ulcerative colitis, Type 1 diabetes, Crohn's disease	Barrett *et al*. (2008); Barrett *et al*. (2009); McGovern *et al*. (2010)
1q31.2[a]	*RGS1*	Crohn's disease	Hunt *et al*. (2008); Parkes *et al*. (2007)
1p31.3[c]	*NFIA*	Ulcerative colitis, Crohn's disease, Psoriasis, Ankylosing spondylitis, Type 1 diabetes	Barrett *et al*. (2009); Nair *et al*. (2009); McGovern *et al*. (2010); Reveille *et al*. (2010)
1p36.11[b]	*RUNX3*	Psoriasis	Nair *et al*. (2009)
1p36.23[c]	*PARK7, TNFRSF9*	Ulcerative colitis, Crohn's disease	Franke *et al*. (2010); Anderson *et al*. (2011)
1p36.32[b]	*TNFRSF14, MMEL1*	Rheumatoid arthritis, Ulcerative colitis	Plenge *et al*. (2010); Anderson *et al*. (2011)
2q12.1[a]	*IL18RAP, IL18R1, IL1RL1, IL1RL2, SLC9A4*	Crohn's disease	Hunt *et al*. (2008); Franke *et al*. (2010)
2p14[b]	*PLEK*	Rheumatoid arthritis	Plenge *et al*. (2010)
2p16.1[a]	*REL, AHSA2*	Ulcerative colitis, Crohn's disease, Psoriasis, Rheumatoid arthritis	Trynka *et al*. (2009); Franke *et al*. (2010); McGovern *et al*. (2010); Plenge *et al*. (2010)
2q31.3[a]	*ITGA4, UBE2E3*	Ankylosing spondylitis	Garner *et al*. (2009); Reveille *et al*. (2010)
2q33.2[a]	*CTLA4, ICOS, CD28*	Type 1 diabetes, Rheumatoid arthritis	Smyth *et al*. (2008); Barrett *et al*. (2009); Plenge *et al*. (2010)
3q13.33[b]	*CD80, KTELC1*	None	
3p14.1[c]	*FRMD4B*	None	
3p21.31[a]	*CCR1, CCR2, CCRL2, CCR3, CCR5, CCR9*	Ulcerative colitis, Crohn's disease	Hunt *et al*. (2008); Daly *et al*. (2008); McGovern *et al*. (2010)
3p22.3[b]	*CCR4*	None	
3q25.33[a]	*IL12A, SCHIP1*	Multiple sclerosis	Hunt *et al*. (2008); De Jager *et al*. (2009)
3q26.2[c]	Intergenic	None	
3q28[a]	*LPP*	None	Hunt *et al*. (2008)
4q27[a]	*IL2, IL21, KIAA1109, TENR, ADAD1*	Ulcerative colitis, Type 1 diabetes, Rheumatoid arthritis	van Heel *et al*. (2007); Barrett *et al*. (2009); Plenge *et al*. (2010); Anderson *et al*. (2011)
6q15[b]	*BACH2, MAP3K7*	Type 1 diabetes, Crohn's disease	Barrett *et al*. (2009); Franke *et al*. (2010)
6q22.33[b]	*PTPRK, THEMIS*	None	

(*Continued*)

Table 1 Continued

Associated coeliac loci	Reported genes	Overlapping autoimmune diseases[d]	References
6q23.3[a]	*TNFAIP3, OLIG3*	Psoriasis, Rheumatoid arthritis, Systemic lupus erythematosus, Ulcerative colitis	Graham *et al*. (2008); Nair *et al*. (2009); Trynka *et al*. (2009); Plenge *et al*. (2010); Anderson *et al*. (2011)
6q25.3[a]	*TAGAP*	Crohn's disease	Hunt *et al*. (2008); Franke *et al*. (2010)
6p25.3[c]	*IRF4*	None	
7p14.1[c]	*ELMO1*	None	
8q24.21[b]	Intergenic	Multiple Sclerosis, Crohn's disease	Bahlo *et al*. (2009); Franke *et al*. (2010)
11q24.3[b]	*ETS1*	Systemic lupus erythematosus	Yang *et al*. (2010)
10q22.3[b]	*ZMIZ1*	Multiple sclerosis, Crohn's disease, Early onset IBD	De Jager *et al*. (2009); Imielinski *et al*. (2009); Franke *et al*. (2010)
12q24.12[a]	*SH2B3, ATXN2*	Rheumatoid arthritis, Type 1 diabetes	Hunt *et al*. (2008); Barrett *et al*. (2009); Plenge *et al*. (2010)
13q14.2[c]	Intergenic	None	
14q24.1[c]	*ZFP36L1*	Type 1 diabetes, Crohn's disease	Barrett *et al*. (2009); Franke *et al*. (2010)
16p13.13[b]	*CIITA, SOCS1, CLEC16A*	Multiple sclerosis, Type 1 diabetes, Ulcerative colitis	Barrett *et al*. (2009); McGovern *et al*. (2010); De Jager *et al*. (2009)
17q21.31[c]	Intergenic	None	
18p11.21[a]	*PTPN2*	Type 1 diabetes, Crohn's disease	Barrett *et al*. (2008); Barrett *et al*. (2009)
21q22.3[b]	*ICOSLG*	Type 1 diabetes, Crohn's disease, Rheumatoid arthritis, Ulcerative colitis	Barrett *et al*. (2008); Smyth *et al*. (2008); Barrett *et al*. (2009); Plenge *et al*. (2010); Anderson *et al*. (2011)
22q11.21[c]	*UBE2L3, YDJC*	Crohn's disease, Systemic lupus erythematosus	Franke *et al*. (2010); Yang *et al*. (2010)
Xp22.2[c]	*TLR7, TLR8*	None	

[a]Loci reported prior to Dubois *et al*. (2010).
[b]Coeliac loci with genome wide significance at $P_{combined} < 5 \times 10^{-8}$.
[c]Coeliac loci with suggestive evidence at either $10^{-6} > P_{combined} > 5 \times 10^{-8}$ or $P_{GWAS} < 10^{-4}$ and $P_{follow-up} < .01$.
[d]Autoimmune diseases with overlapping loci according to the Catalogue of Published Genome-Wide Association Studies (http://www.genome.gov/26525384); accessed on 31 March 2011.

genes constitute approximately 1% of the genome but harbour 85% of mutations with large effects on disease traits and an excess of low frequency nonsynonymous variants (Li *et al*., 2010). It is largely acknowledged that mutations found in the exome are more rare as these regions influence expression of proteins, leading to deleterious effects, so are prevented from attaining a high frequency by selection. Present discoveries have substantiated these observations (Zhu *et al*., 2011). There has been a dramatic surge in exome capture and massively parallel sequencing in rare disease, which have historically been constrained by small kindred sizes and locus heterogeneity. Ng *et al*. (2009) published the first proof of principle paper (Ng *et al*., 2009).

The obvious setback to exome sequencing is information on only the coding region of the genome is analysed. To truly elucidate the relationship between disease phenotype and their corresponding genetic basis, attention ought to be focused on the entire genome. With ongoing decrease in sequencing cost, this technique may be affordable in a large

cohort of patients required for complex disease genetics. **See also**: Next Generation Sequencing Technologies and Their Applications

Concluding remarks

This review has discussed the genetics of CD, composed of a strong HLA background and current associated loci found through GWAS, and the next steps for casual variant discovery in genetics. With any complex disease, it is important to appreciate types of variation in the human genome and their effects on each other. It is unlikely that common and rare variants in only one gene are common to all autoimmune diseases (Surolia *et al*., 2010), as it is evident multiple genes show interaction in multiple immune pathways. Both rare and common variants are thought to be attributable to disease onset, but gene–gene and gene–environment interactions should be taken into account when construing biological disease pathways. For complete resolution of the CD, exome sequencing is the next step for rare functional discovery, as well as deep resequencing of associated genes in cases and controls. For association-based studies, Immunochip and content incorporated onto arrays from sequencing projects, such as 1000G and UK10K (www.uk10k.org), will drive a new generation of GWAS focusing on rare variation, supported by functional studies on candidate genes of interest.

References

Abadie V, Sollid LM, Barreiro LB and Jabri B (2011) Integration of genetic and immunological insights into a model of celiac disease pathogenesis. *Annual Reviews in Immunology* **29**: 493–525.

Al-toma A, Visser OJ, van Roessel HM *et al*. (2007) Autologous hematopoietic stem cell transplantation in refractory celiac disease with aberrant T cells. *Blood* **109**(5): 2243–2249.

Anderson CA, Boucher G, Lees CW *et al*. (2011) Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nature Genetics* **43**(3): 246–252.

Babron MC, Nilsson S, Adamovic S *et al*. (2003) Meta and pooled analysis of European coeliac disease data. *European Journal of Human Genetics* **11**(11): 828–834.

Bahlo M, Rubio JP and Booth DR (2009) Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20. *Nature Genetics* **41**(7): 824–828.

Barrett JC, Clayton DG, Concannon P *et al*. (2009) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature Genetics* **41**(6): 703–707.

Barrett JC, Hansoul S, Nicolac DL *et al*. (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature Genetics* **40**(8): 955–962.

van Belzen MJ, Koeleman BP, Crusius JB *et al*. (2004) Defining the contribution of the HLA region to cis DQ2-positive coeliac disease patients. *Genes & Immunity* **5**(3): 215–220.

van Belzen MJ, Meijer JWR, Sandkuijl LA *et al*. (2003) A major non-HLA locus in celiac disease maps to chromosome 19. *Gastroenterology* **125**(4): 1032–1041.

Brophy K, Ryan AW, Thornton JM *et al*. (2006) Haplotypes in the CTLA4 region are associated with coeliac disease in the Irish population. *Genes & Immunity* **7**(1): 19–26.

Brown WM, Pierce J, Hilner JE *et al*. (2009) Overview of the MHC fine mapping data. *Diabetes Obesity & Metabolism* **11**(1): 2–7.

Caruso R, Fina D, Peluso I *et al*. (2007) A functional role for interleukin-21 in promoting the synthesis of the T-cell chemoattractant, MIP-3alpha, by gut epithelial cells. *Gastroenterology* **132**(1): 166–175.

Catassi C, Ratsch IM, Gandolfi L *et al*. (1999) Why is coeliac disease endemic in the people of the Sahara? *Lancet* **354**(9179): 647–648.

Chen R, Stahl EA, Kurreeman FA *et al*. (2011) Fine mapping the TAGAP risk locus in rheumatoid arthritis. *Genes & Immunity* **12**(4): 314–318.

Cummins AG and Roberts-Thomson IC (2009) Prevalence of celiac disease in the Asia-Pacific region. *Journal of Gastroenterology and Hepatology* **24**(8): 1347–1351.

Daly MJ, Barrett JC, Hansoul S *et al*. (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature Genetics* **40**(8): 955–962.

De Jager PL, Jia X, Wang J *et al*. (2009) Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. *Nature Genetics* **41**(7): 776–782.

Dieterich W, Laag E, Schopper H *et al*. (1998) Autoantibodies to tissue transglutaminase as predictors of celiac disease. *Gastroenterology* **115**(6): 1317–1321.

Dube C, Rostom A, Sy R *et al*. (2005) The prevalence of celiac disease in average-risk and at-risk Western European populations: a systematic review. *Gastroenterology* **128**(4 suppl. 1): 57–67.

Dubois PC and van Heel DA (2008) Translational mini-review series on the immunogenetics of gut disease: immunogenetics of coeliac disease. *Clinical & Experimental Immunology* **153**(2): 162–173.

Dubois PC, Trynka G, Franke L *et al*. (2010) Multiple common variants for celiac disease influencing immune gene expression. *Nature Genetics* **42**(4): 295–302.

Durbin RM, Abecasis GR, Altshuler DL *et al*. (2010) A map of human genome variation from population-scale sequencing. *Nature* **467**(7319): 1061–1073.

Fontenot JD, Rasmussen JP, Gavin MA, and Rudensky AY. (2005) A function for interleukin 2 in Foxp3-expressing regulatory T cells. *Nature Immunology* **6**(11): 1142–1151.

Franke A, Parkes M, McGovern DPB *et al*. (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genetics* **42**(12): 1118–1118.

Garner CP, Murray JA, Ding YC *et al*. (2009) Replication of celiac disease UK genome-wide association study results in a US population. *Human Molecular Genetics* **18**(21): 4219–4225.

Graham RR, Cotsapas C, Davies L *et al*. (2008) Genetic variants near TNFAIP3 on 6q23 are associated with systemic lupus erythematosus. *Nature Genetics* **40**(9): 1059–1061.

Greco L, Romino R, Coto I *et al*. (2002) The first large population based twin study of coeliac disease. *Gut* **50**(5): 624–628.

Gutierrez-Achury J, de Almeida RC, Wijmenga C *et al*. (2011) Shared genetics in celiac disease and other immune-mediated diseases. *Journal of Internal Medicine* **269**(6): 591–603.

Han SB, Moratz C, Huang NN *et al*. (2005) Rgs1 and Gnai2 regulate the entrance of B lymphocytes into lymph nodes and B

cell motility within lymph node follicles. *Immunity* **22**(3): 343–354.

Heap GA and van Heel DA (2009) Genetics and pathogenesis of coeliac disease. *Seminars in Immunology* **21**(6): 346–354.

van Heel DA, Franke L, Hunt KA *et al.* (2007) A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. *Nature Genetics* **39**(7): 827–829.

Holtmann MH and Neurath MF (2004) T helper cell polarisation in coeliac disease: any (T-)bet? *Gut* **53**(8): 1065–1067.

Hughes T, Kim-Howard X, Kelly JA *et al.* (2011) Fine mapping and trans-ethnic genotyping establish IL2/IL21 genetic association with lupus and localize this genetic effect to IL21. *Arthritis & Rheumatism* **63**(6): 1689–1697.

Hugot JP, Chamaillard M, Zouali H *et al.* (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**(6837): 599–603.

Hunt KA, Zhernakova A, Turner G *et al.* (2008) Newly identified genetic risk variants for celiac disease related to the immune response. *Nature Genetics* **40**(4): 395–402.

Hunt KA, McGovern DP, Kumar PJ *et al.* (2005) A common CTLA4 haplotype associated with coeliac disease. *European Journal of Human Genetics* **13**(4): 440–444.

Imielinski M, Hakonarson H, Baldassano RN *et al.* (2009) Common variants at five new loci associated with early onset inflammatory bowel disease. *Nature Genetics* **41**(12): 1335–1340.

Karell K, Louka AS, Moodie SJ *et al.* (2003) HLA types in celiac disease patients not carrying the DQA1*05-DQB1*02 (DQ2) heterodimer: results from the European Genetics Cluster on Celiac Disease. *Human Immunology* **64**(4): 469–477.

Kaukinen K, Partanen J, Maki M and Collin P (2002) HLA-DQ typing in the diagnosis of celiac disease. *American Journal of Gastroenterology* **97**(3): 695–699.

King AL, Moodie SJ, Fraser JS *et al.* (2003) Coeliac disease: investigation of proposed causal variants in the CTLA4 gene region. *European Journal of Immunogenetics* **30**(6): 427–432.

King AL, Yiannakou JY, Brett PM *et al.* (2000) A genome-wide family based linkage study of coeliac disease. *Annals of Human Genetics* **64**(part 6): 479–490.

Li Y, He X, Schembri-King J, Jakes S and Hayashi J. (2000) Cloning and characterization of human Lnk, an adaptor protein with pleckstrin homology and Src homology 2 domains that can inhibit T cell activation. *Journal of Immunology* **164**(10): 5199–5206.

Li Y, Vinckenbosch N, Tian G *et al.* (2010) Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nature Genetics* **42**(11): 969–972.

Lowe CE, Cooper JD, Brusko T *et al.* (2007) Large-scale genetic fine mapping and genotype–phenotype associations implicate polymorphism in the IL2RA region in type 1 diabetes. *Nature Genetics* **39**(9): 1074–1082.

Maiti AK, Kim-Howard X, Viswanathan P *et al.* (2010) Confirmation of an association between rs6822844 at the Il2-Il21 region and multiple autoimmune diseases: evidence of a general susceptibility locus. *Arthritis & Rheumatism* **62**(2): 323–329.

McGovern DP, Gardet A, Torkvist L *et al.* (2010) Genome-wide association identifies multiple ulcerative colitis susceptibility loci. *Nature Genetics* **42**(4): 332–337.

Molberg O, Mcadam SN, Korner R *et al.* (1998) Tissue transglutaminase selectively modifies gliadin peptides that are recognized by gut-derived T cells in celiac disease (vol 4, pg 713, 1998). *Nature Medicine* **4**(8): 974–974.

Momozawa Y, Mni M, Nakamura K *et al.* (2011) Resequencing of positional candidates identifies low frequency IL23R coding variants protecting against inflammatory bowel disease. *Nature Genetics* **43**(1): 43–47.

Monsuur AJ, de Bakker PI, Zhernakova A *et al.* (2008) Effective detection of human leukocyte antigen risk alleles in celiac disease using tag single nucleotide polymorphisms. *PLoS One* **3**(5): e2270.

Nair RP, Duffin KC, Helms C *et al.* (2009) Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. *Nature Genetics* **41**(2): 199–204.

Nejentsev S, Walker N , Riches D, Egholm M and Todd JA (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* **324**(5925): 387–389.

Ng PC, Levy S, Huang J *et al.* (2008) Genetic variation in an individual human exome. *PLoS Genetics* **4**(8): e1000160.

Ng SB, Turner EH, Robertson PD *et al.* (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**(7261): 272–276.

Parkes M, Barrett JC, Prescott NJ *et al.* (2007) Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nature Genetics* **39**(7): 830–832.

Percopo S, Babron MC, Whalen M *et al.* (2003) Saturation of the 5q31–q33 candidate region for coeliac disease. *Annals of Human Genetics* **67**(part 3): 265–268.

Plenge RM, Stahl EA, Raychaudhuri S *et al.* (2010) Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nature Genetics* **42**(6): 508–U556.

Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? *American Journal of Human Genetics* **69**(1): 124–137.

Reveille JD, Sims AM, Danoy P *et al.* (2010) Genome-wide association study of ankylosing spondylitis identifies non-MHC susceptibility loci. *Nature Genetics* **42**(2): 123–127.

Smyth DJ, Plagnol V, Walker NM *et al.* (2008) Shared and distinct genetic variants in type 1 diabetes and celiac disease. *New England Journal of Medicine* **359**(26): 2767–2777.

Sarra M, Franze E, Pallone F and Monteleone G (2011) Targeting interleukin-21 in inflammatory diseases. *Expert Opinion on Therapeutic Targets* **15**(6): 695–702.

Sollid LM, Markussen G, Ek J *et al.* (1989) Evidence for a primary association of celiac disease to a particular HLA-DQ alpha/beta heterodimer. *Journal of Experimental Medicine* **169**(1): 345–350.

Strengell M, Sareneva T, Foster D, Julkunen I and Matikainen S (2002) IL-21 up-regulates the expression of genes associated with innate immunity and Th1 response. *Journal of Immunology* **169**(7): 3600–3605.

Surolia I, Pirnie SP, Chellappa V *et al.* (2010) Functionally defective germline variants of sialic acid acetylesterase in autoimmunity. *Nature* **466**(7303): 243–247.

Tosi R, Vismara D, Tanigaki N *et al.* (1983) Evidence that celiac disease is primarily associated with a DC locus allelic specificity. *Clinical Immunology and Immunopathology* **28**(3): 395–404.

Tran T, Paz P, Velichko S *et al.* (2010) Interferonbeta-1b induces the expression of RGS1 a negative regulator of G-protein signaling. *International Journal of Cell Biology* **2010**: 529376.

Trynka G, Zhernakova A, Romanos J *et al*. (2009) Coeliac disease-associated risk variants in TNFAIP3 and REL implicate altered NF-kappaB signalling. *Gut* **58**(8): 1078–1083.

Yang WL, Shen N, Ye DQ *et al*. (2010) Genome-wide association study in Asian populations identifies variants in ETS1 and WDFY4 associated with systemic lupus erythematosus. *Plos Genetics* **6**(2): e1000841.

Zhernakova A, Elbers CC, Ferwerda B *et al*. (2010) Evolutionary and functional analysis of celiac risk loci reveals SH2B3 as a protective factor against bacterial infection. *American Journal of Human Genetics* **86**(6): 970–977.

Zhu Q, Ge D, Maia JM *et al*. (2011) A genome-wide comparison of the functional properties of rare and common genetic variants in humans. *American Journal of Human Genetics* **88**(4): 458–468.

Further Reading

Anderson RP (2008) Coeliac disease: current approach and future prospects. *Internal Medicine Journal* **38**: 790–799.

Baranzini SE (2009) The genetics of autoimmune diseases: a networked perspective. *Current Opinion in Immunology* **21**(6): 596–605.

Deiterich W, Ehnis T, Bauer M *et al*. (1997) Identification of tissue transglutaminase as the autoantigen of celiac disease. *Nature Medicine* **3**: 797–801.

Eichler EE, Flint J , Gibson G *et al*. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics* **11**: 446–450.

Knight JC (2009) *Human Genetic Diversity Functional Consequences for Health and Disease*. Oxford: Oxford University Press.

Manolio TA, Collins FS , Cox NJ *et al*. (2009) Finding the missing heritability of complex diseases. *Nature* **461**: 747–753.

Marsh MN (1992) Gluten, major histocompatibility complex and the small intestine: a molecular and immunobiologic approach to the spectrum of gluten sensitivity ('celiac sprue'). *Gastroenterology* **102**(1): 330–354.

Sollid LM (2000) Molecular basis of celiac disease. *Annual Reviews in Immunology* **18**: 53–81.

Van Heel DA and Hunt KA (2005) Genetics in coeliac disease. *Best Practice & Research: Clinical Gastroenterology* **19**(3): 323–339.